BAJOPAS Volume 14 Number 2, December, 2021 **http://dx.doi.org/10.4314/bajopas.v14i2.16**

Bayero Journal of Pure and Applied Sciences, 14(2): 134 - 149
Received: August, 2021
Accepted: September, 2021
ISSN 2006 – 6996

COMPARING THE PREDICTION ACCURACY OF RIDGE, LASSO AND ELASTIC NET REGRESSION MODELS WITH LINEAR REGRESSION USING BREAST CANCER DATA

[1]Usman, M., [1]Doguwa S. I. S. and [2]Alhaji, B. B.
[1]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria
[2]Department of Mathematics, Nigerian Defense Academy, Kaduna, Nigeria
Corresponding Author: Email and Phone (ummohammed67@yahoo.com, 08035940877)

ABSTRACT
Regularised regression methods have been developed in the past to overcome the shortcomings of ordinarily least squares (OLS) regression of not performing well with respect to both prediction accuracy and model complexity. OLS method may fail or produce regression estimates with high variance in the presence of multi-collinearity or when the predictor variables are greater than the number of observations. This study compares the predictive performance and additional information gained of Ridge, Lasso and Elastic net regularised methods with the classical OLS method using data of breast cancer patients. The findings have shown that using all the predictor variables, the OLS method failed because of the presence of multiple collinearity, while regularised Ridge, Lasso and Elastic net methods produced results that showed the predictor variables mostly significant. Using the training data, the Elastic net and Lasso seemed to indicate more significant predictor variables than the Ridge method. The result also indicated that breast cancer patients in age groups 30-39, those that are married and in stage1 of the disease, have longer survival times, while patients that are in stage2 and stage3 have shorter survival times. The OLS regression produced results only when four of the predictor variables were excluded; even then, the regularised methods still outperformed the OLS regression in terms of prediction accuracy.

Keywords: OLS, Ridge, Lasso, Elastic Net, Breast Cancer Data

INTRODUCTION

Breast cancer is the most frequent cancer in women and is the second most common cancer across the globe. In 2018, it was responsible for an estimated 2.1 million cancers, accounting for the fifth leading cause of cancer deaths worldwide (Bray *et al* 2018, Fitzmaurice, 2018). One in every 9 women in developed countries and one in every 20 women in less developed nations may have the risk of breast cancer (Fitzmaurice, 2018).

Massive amount of data with increasing dimensions are being generated in many areas of life, such as medicine, economics, social sciences etc. The massive amount of data is in two dimensions which include the dependent and predictor variables and the number of observations.

In biological data there are often fewer observations available than predictor variables. For instance, gene expression data include more than ten thousand gene profiles from hundreds of patients (see Shen *et al,* 2011). Variables and features selection methods have become the focus of research in areas of application for which large datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents (see Talib *et al,* 2016), gene expression array analysis and combinatorial chemistry (Liv *et al,* 2017). The reasons for variable selection are three-fold: improving the prediction performance of the exposure variables; providing faster and more cost-effective predictors; and providing a better understanding of the underlying process that generated the data (Guyon & Elisseeff, 2003). However, it is undesirable to keep irrelevant predictors in the final model since this makes it difficult to interpret the resultant model and may decrease its predictive ability. In the regularization framework, many different types of penalties have been introduced to achieve variable selection (Yichao & Yufeng, 2009). Ordinary least squares (OLS) method is a popular procedure in regression methods. It is expected not to perform well with respect to both prediction accuracy and model complexity.

OLS regression may result in higher estimates of the regression coefficients in the presence of collinearity or when the number of predictor variables (P) is large relative to the number of observation (N) (VanderKooij, and Meulman 2006, Wessel and VanWieringen, 2020). The objective of this paper is to analyze the breast cancer data, sourced from Ahmadu Bello University Teaching Hospital, Zaria-Nigeria and to find out which of the predictor variables exert greater influence on the survival of the breast cancer patients. Secondly, the exposure variables shall be used to predict the breast cancer patient's survival times and compare the prediction accuracy of Ridge, Lasso and Elastic net as regularised models with the classical linear regression model, using matrix algebra.

MATERIALS AND METHODS

Ordinary Least Squares (OLS)

In multiple linear regression, we consider the following relationship between predictor variables X_1, X_2, ……. X_p and the response variable Y i.e

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots\ldots + \beta_p X_p + \xi \qquad (1)$$

where ξ is the n column vector of random observation or error term and β_0 , β_1 , β_2 ,………….β_p are the regression coefficients. The objective is to find the estimated regression model

$$\hat{y}_i = b_o + b_1 X_{i1} + b_2 X_{i2} \ldots\ldots\ldots\ldots b_p X_{ip} \qquad (2)$$

where X_{ij} is the jth predictor variable observation on the i th subject with i = 1, 2, ……. n and j = 0, 1, 2, ……. p . The procedure and computation involved in the linear multiple regression analysis remain the same for any number of predictor variables. Equation (1) can be expressed in matrix form (Melkumova and Shatskikh, 2017).

$$Y = X\beta + \xi \qquad (3)$$

where **Y** is an n column vector of dependent variable; **X** is n x p matrix of predictor variables and β is a p parameter vector. The n column vector **ξ** is vector of error terms. Also, b_0, b_1, b_2,……….,b_p are the estimators of the unknown vector **β,** The objective of the regression analysis is to estimate the vector β based on the X and Y observations:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} & x_{1p} \\ 1 & x_{21} & & x_{2,p-1} & x_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x_{n-1,1} & \cdots & x_{n-1,p-1} & x_{n-1,p} \\ 1 & x_{n1} & & x_{n,p-1} & x_{n,p} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$$

And $\xi_i = y_i - \sum_{j=0}^{p} \beta_j x_{ij}$, i = 1, 2,…., n .

The Ordinarily Least Squares (OLS) estimator is used to estimate the parameters where

$$\xi^t \xi = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2$$

is minimized. Given that det (X^TX) > 0, the OLS estimates can be obtained in matrix form as

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (4)$$

and variance of $\hat{\beta}$ is

$$\hat{Var}(\hat{\beta}) = \sigma^2 \, \mathrm{diag}[(X^T X)^{-1}] \qquad (5)$$

Ridge regression

In high-dimensional settings when the number of covariates p exceeds the number of observations (n), the problem of maximizing the partial log-likelihood cannot be done uniquely. A way to deal with the $p >> n$ situation is to introduce a penalty term into the partial log-likelihood $l(\beta)$, referred to as regularization. This approach is also reasonable when covariates are less than the observations ($p < n$) settings since it considers collinearity among the predictors and helps to prevent over-fitting (Madjar, 2018). Ridge regression proposed by Hoerl & Kennard (1970) is one of the penalization or regularization methods that reduce this variability by shrinking the coefficients, resulting in more prediction accuracy at the cost of a small increase of bias. In Ridge regression, the coefficients are shrunken towards zero, but will never become exactly zero. So, when the number of predictors is large, Ridge regression will not provide a sparse model that is easy to interpret. OLS estimate depends on $(X^TX)^{-1}$ where if the rank(X) is less than the number of predictors (P) then (X^TX) will not have an inverse. In this situation, Ridge regression can overcome this problem by constraining the coefficient estimates; and thus reduce the estimator's variance and introduce some bias. Rebecca *et al.* (2015) estimated the parameters using Ridge regression method that exhibits the least bias on large data sets in their study on penalized likelihood methods that improve parameter estimates in occupancy models. With the problem of multi-collinearity, Ridge regression improves the prediction performance.

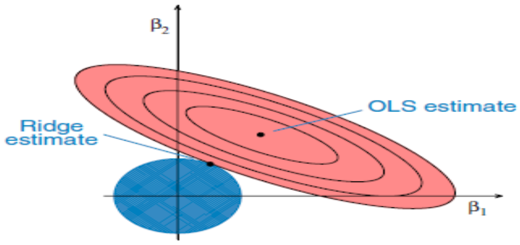

Fig 1: Graphical representation of Ridge Regression and OLS
(see https://onlinecourses.science.psu. edu/stat857/ node/155/.)

The Ordinarily Least Square Estimates are unbiased, but produce large prediction variance. Therefore; to improve the accuracy of the prediction is to either shrink the values of the regression coefficients toward zero or by setting some insignificant coefficients to zero. Thus, the accuracy of the prediction will be improved. This can be done by introducing some estimation bias or constraint and the variance can be reduced, and this can result in reducing the mean squared error of prediction. Ridge regression is a popular method in the context of multi-collinearity. Therefore, Ridge regression imposes a constraint on the coefficients and the coefficients are estimated by minimizing the penalized sum of squares.

From Figure 1, the least squares solution is the centre of the ellipse i.e OLS estimate. The ellipse that is centered around the OLS estimate represents the region of constant Residual Sums of Squares (RSS). Ridge regression has a blue circular constraint with no sharp points, the intersection between the red ellipse and the blue circle will not generally happen on the x – axis and therefore the Ridge regression coefficient estimates will be exclusively non-zero. The ellipse corresponds to the contours of the RSS; the inner ellipse has smaller RSS, and RSS is minimised at ordinarily least square estimates. Ridge regression minimises the residual sum of squares with the sum of square value of the coefficients. Ridge regression model can be formulated as follows:

$$\hat{\beta}_{ridge} = \arg\min\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta^2{}_j\right] \qquad (6)$$

where n is the number of observations, $\lambda > 0$ is a tuning parameter which is the amount of shrinkage of the coefficients and p is the number of predictors. When $\lambda = 0$, equation (6), corresponds to least squares regression. Now for real valued function f, with domain S, arg min [f (β)] $\in$ Sf(β) is the set of element in S that achieve the global minimum in S. Using matrix algebra, we can write equation (5) in matrix form as:

$$\overset{\wedge}{\beta}_{(ridge)} = (X^{T}X + \lambda I)^{-1} X^{T}Y \quad (7)$$

where **I** is the p x p identity matrix. Adding λ to the diagonal of **X^T^X** makes the problem nonsingular even with the multi-collinearity in the data. So, we compute equation (7) for a range of λ values (λ = 0.01, 0.02, 0.03, ….,1) say and choose the optimal $\overset{\wedge}{\beta}_{(ridge)}$ that minimises the mean squared error MSE(λ),

$$MSE(\lambda) = (Y - \widehat{Y}_{(\lambda)})^{T}(Y - \widehat{Y}_{(\lambda)}) / \mathrm{n} \quad (8)$$

where

$$\widehat{Y}_{(\lambda)} = X\widehat{\beta}_{(ridge)} \quad (9)$$

Our optimal Ridge regularized parameter estimate $optimal\widehat{\beta}_{(Ridge)}$ is then given as,

$$optimal\hat{\beta}_{(ridge)} = (X^{T}X + \lambda_{opt}I)^{-1} X^{T}Y \quad (10)$$

where λ_{opt} is the value of λ in which MSE(λ) attains the global minimum. The variance of the estimate is given by

$$Var(\hat{\beta}_{ridge}) = \sigma^{2}diag[(X^{T}X + \lambda I)^{-1}X^{T}X(X^{T}X + \lambda I)^{-1}] \quad (11)$$

Least Absolute Shrinkage and Selection Operator (Lasso)

Tibshirani (1996) proposed a method called Least Absolute Shrinkage and Selection Operator (Lasso), similar to Ridge regression in dealing with many predictor variables. It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical regression model (Emmeet-Streib and Dehmer, 2019). However, Lasso is different from Ridge regression because it deals with variable selection and shrinkage of the parameter. Lasso minimises the residual sum of squares subject to the sum of the absolute values of the coefficients. Because of the constraint, Lasso method shrinks some regression coefficients toward zero and others to exactly zero and hence produces a sparse model. Lasso is a method of selecting a subset of variables in a model while simultaneously shrinking the other regression coefficients toward zero, due to some constraints in Lasso principles. The popularity of the classical Lasso lies in its ability to shrink coefficients to zero, thereby automatically performing variable selection, and the effect of the penalization is that Lasso sets the $\overset{\wedge}{\beta}_{j}$ s for some variables to zero. In other words, it does the model selection for us (Van Erp *et al,* 2019, Ahrens *et al* 2018, Chaturvedi, 2018). Goeman (2010), efficiently computes estimates of parameters in high dimensional model using L$_1$ penalized (lasso) method. The dimensionality of the collected data in clinical studies for complex disease such as cancer for example, is growing exceedingly fast. It is analytically challenging for researchers to elucidate the relationship between the most influential factors (variables) and patient survival outcomes (Xiao *et al.*, 2016). Tuji (2010) applied L$_1$ regularization form of Lasso to select the most significant variables on the survival dataset in their cancer study. Lasso sets coefficients to zero exactly if the variables are not important when (λ) is large enough. Ridge regression and Lasso minimise the RSS with the penalty term as constraints which means that the shrinkage problem will find the smallest RSS within a budget defined by:

a) a circle for Ridge regression
b) a diamond for Lasso (absolute value). The absolute values are going to be a constraint region that has sharp corners.

The solution will be, the first place the RSS contours hit the constraint region. In high dimensions with Lasso, you have edges and corners that make the diamond, and along an edge or a corner, if you hit there, you get a zero. So this is, geometrically, why you get sparsity in the Lasso (https://datacadamia.com/data mining/lasso).

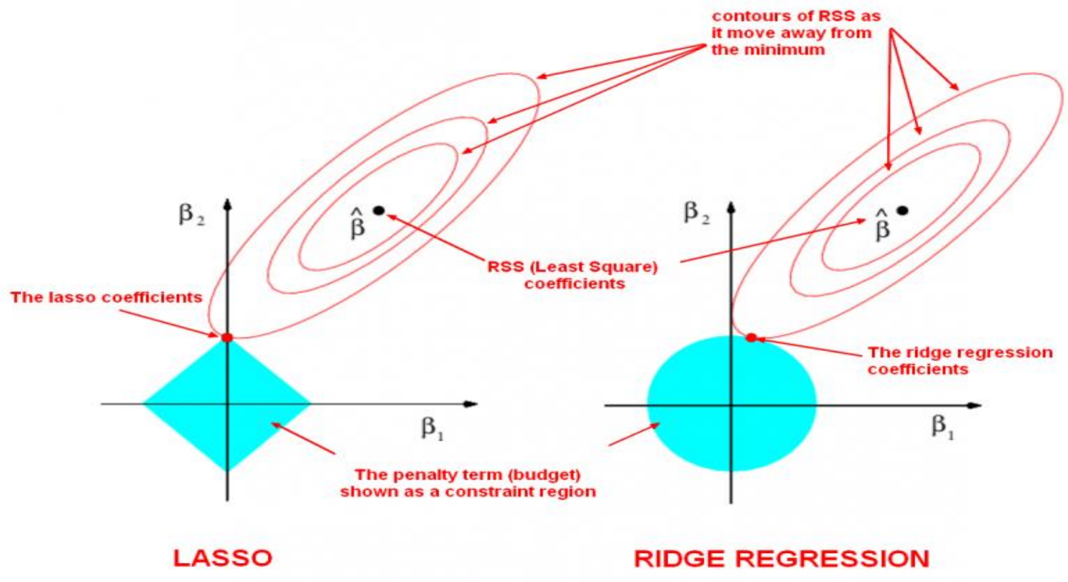
contours of RSS as it move away from the minimum
β_2
$\hat{\beta}$
RSS (Least Square) coefficients
The lasso coefficients
β_1
The penalty term (budget) shown as a constraint region
β_2
$\hat{\beta}$
The ridge regression coefficients
β_1
LASSO
RIDGE REGRESSION

Fig 2: Lasso Regression (left) vs. Ridge Regression (right)
(https://datacadamia.com/data_mining/lasso)

The Lasso penalty regularised the linear regression coefficients of penalized least square criterion as:

$$\hat{\beta}_{(\text{lasso})} = \arg\min\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right] \tag{12}$$

where n is the number of observations, p is the number of predictors for example, genes. λ is the tuning parameter which determines the amount of shrinkage of the regression coefficients. The higher the value of λ, the greater will be the shrinkage of the β coefficient as seen in equation (8) and this in turn, will make the coefficients more robust to collinearity.

Lasso performed better than Ridge in scenarios with many noise predictors and worse in the presence of correlated predictors (Pavlou *et al,* 2015).

To obtain the $\boldsymbol{\beta}_{lasso}$ in matrix form, we need to minimise equation (13) with respect to **β:**

$$f(\beta) = \left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right] \tag{13}$$

This involves differentiating the equation with respect to **β** and setting the derivative to zero to in order to obtain the system of equations:

$$(X^{T}X)\beta + 0.5\lambda\beta^{*} = X^{T}Y \tag{14}$$

where $\boldsymbol{\beta}^{*}$ is defined as,

$$\boldsymbol{\beta}^{*} = \begin{pmatrix} \beta_0 / |\beta_0| \\ \beta_1 / |\beta_1| \\ . \\ . \\ \beta_p / |\beta_p| \end{pmatrix} \tag{15}$$

Clearly equation (14) is not in closed form, so iterative method has to be used to determine the lasso estimate of **β.** Using the Newton Raphson Algorithm, and for a given λ > 0 and initial value of $\boldsymbol{\beta_0}$ one can run the iteration in equation (11):

$$\boldsymbol{\beta_t} = \boldsymbol{\beta}_{t-1} - (X^{T}X + 0.5\lambda\boldsymbol{G}_{t-1})^{-1}\,(X^{T}X\boldsymbol{\beta}_{t-1} + 0.5\lambda\boldsymbol{\beta}^{*}_{t-1} - X^{T}Y) \tag{16}$$

where the p +1 by p +1 diagonal matrix **G** is defined as,

$$\boldsymbol{G} = diag\left(\left(\frac{1-\beta_0^{-2}}{|\beta_0|}\right), \left(\frac{1-\beta_1^{-2}}{|\beta_1|}\right), \left(\frac{1-\beta_2^{-2}}{|\beta_2|}\right), \ldots.., \left(\frac{1-\beta_p^{-2}}{|\beta_p|}\right)\right) \tag{17}$$

and t =1, 2, 3, …. , until convergence is achieved. A possible convergence criterion could be to stop the iteration whenever:

$$(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^T(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})/(\boldsymbol{\beta}_{t-1})^T(\boldsymbol{\beta}_{t-1}) < 10^{-6} \qquad (18)$$

and take $\widehat{\boldsymbol{\beta}}_{lasso} = \boldsymbol{\beta}_{t-1}$ for the given λ.

To obtain our optimal $\widehat{\boldsymbol{\beta}}_{lasso}$, we run equation (16) for a range of λ values (λ=0.01, 0.02, 0.03, ….,1) say and choose the optimal $\hat{\beta}_{(lasso)}$ that minimises the mean squared error MSE(λ), while the corresponding value of λ gives the λ$_{opt}$

$$MSE(\lambda) = (Y - \widehat{Y}_{(\lambda)})^T (Y - \widehat{Y}_{(\lambda)}) / \mathrm{n} \qquad (19)$$

and

$$\widehat{Y}_{(\lambda)} = X\widehat{\beta}_{(lasso)} \qquad (20)$$

where λ_{opt} is the value of λ in which MSE(λ) attains the global minimum. Alternatively, re-run the iteration in equation (16) with λ$_{opt}$ to obtain the optimal $\widehat{\boldsymbol{\beta}}_{lasso}$. The variance of $\widehat{\boldsymbol{\beta}}_{lasso}$ is given by,

$$Var(\hat{\beta}_{lasso}) = \sigma^2 diag[(X^TX + 0.5\lambda G)^{-1}X^TX(X^TX + 0.5\lambda G)^{-1}] \qquad (21)$$

where

$$\sigma^2 = (Y - \widehat{Y}_{(\lambda)})^T (Y - \widehat{Y}_{(\lambda)}) / (\mathrm{n} - \mathrm{p}) \qquad (22)$$

and n is the number of subjects and p is the number of predictor variables.

Elastic net

Elastic net was introduced by Zou and Hastie (2005), to extend the Lasso by improving some of its limitations, especially with respect to the variable selection. The method produces a regression model that is penalized with both the Ridge regression penalty term of L$_1$ - norm and Lasso regression penalty term of L$_2$ - norm. The consequence of this is to effectively shrink coefficients just like in Ridge regression and to set some coefficients to zero as in Lasso. The L$_1$ - norm part of the penalty generates a sparse model by shrinking some regression coefficients exactly to zero. The L$_2$ - norm part of the penalty removes the limitation on the number of selected variables, encourages grouping effect, and stabilizes the L$_1$ regularization path (Park and Konishi, 2015).

In this situation, Elastic net not only selects variables, but may also perform better than Lasso with observations that are collinear. Liu and Li (2017) used an efficient Elastic net with regression coefficients method to select the significant variables of the spectrum data. Steele *et al*, (2018) analyzed the electronic patient health records for predicting patient mortality with Elastic net method. In finance, Ho *et, al* (2015) used Elastic net to define portfolios of stocks and predict the credit ratings of corporations. Furthermore, Elastic net is particularly useful in cases where the number of predictor variables (p) in datasets are much larger than the number of observations (n). In such cases, Lasso is not capable of selecting more than 'n' predictors but the Elastic net has this capability (Frank and Matthias, 2019). Elastic net minimises the loss function and the estimated parameter vector is given by

$$\hat{\beta}_{Elastic} = \arg\min\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\left[(1-\alpha)\sum_{j=1}^{p}\beta_j{}^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right]\right] \qquad (23)$$

where

λ = tuning parameter

α = weight that determines how much should be given to Lasso or Ridge regression,

such that 0 ≤ α ≤ 1, where α = 0, $\hat{\beta}_E$ becomes $\hat{\beta}_R$ and where α = 1 , $\hat{\beta}_E$ becomes $\hat{\beta}_L$

The matrix form (23) does not have a closed form and its solution can only be obtained iteratively using:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - (X^TX + \lambda\{\alpha I + 0.5(1-\alpha)\boldsymbol{G}_{t-1}\})^{-1}(X^TX\boldsymbol{\beta}_{t-1} + \lambda\{\alpha\boldsymbol{\beta}_{t-1} + 0.5(1-\alpha)\boldsymbol{\beta}^*_{t-1}\} - X^TY) \qquad (24)$$

where t = 1, 2, …… and the diagonal matrix **G**, **β** and **β*** are as defined in equation (16). For a given α such that $0 < \alpha < 1$ and a given initial value $\boldsymbol{\beta_0}$, one runs equation (24) for each value of λ (λ=0.01, 0.02, …..,1) say, until convergence is achieved based on the convergence criterion defined in equation (18), and then compute the MSE(λ). The λ_{opt} is obtained as that value of λ that returns the minimum MSE(λ). Having obtained λ_{opt} we can use it to re-run equation (24) with values of α say (α =0.01, 0.02,……..0.99) and for each α we estimate MSE(α).

The optimal α is determined as that α that returns the smallest MSE(α). Both the optimal α and λ are then used in equation (24) to re-run the iteration until convergence. The converged $\widehat{\boldsymbol{\beta}}$ is the elastic net estimate $\widehat{\boldsymbol{\beta}}_{elastic}$ of **β.** The variance of the estimate is then computed as,

$$Var(\hat{\beta}_{elastic}) = \sigma^2\, diag[(X^TX + \lambda\{\alpha I + 0.5(1-\alpha)\boldsymbol{G}_{t-1}\})^{-1}X^TX(X^TX + \lambda\{\alpha I + 0.5(1-\alpha)\boldsymbol{G}_{t-1}\})^{-1}] \qquad (25)$$

where σ^2 is defined as

$$\sigma^2 = (Y - \widehat{Y}_{(\lambda)})^T(Y - \widehat{Y}_{(\lambda)})/(\text{n}-\text{p}) \qquad (26)$$

and

$$\widehat{Y}_{(\lambda)} = X\widehat{\beta}_{(Elastic)} \qquad (27)$$

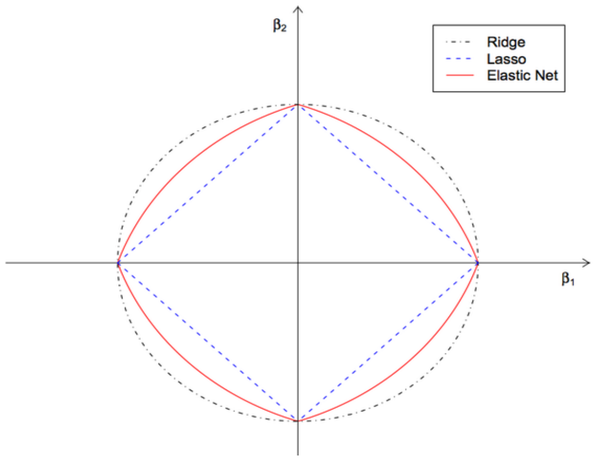

Fig 3: Elastic net vs. Lasso vs. Ridge regression (Sosnovshchenko, nd).

The Elastic net penalty is a convex combination of the Lasso and the Ridge constraint functions. Figure 3 shows the effect of weight α on the regularization. From Fig 3, the Elastic net penalty (in red color or solid line) is located between the Lasso and the Ridge penalties. In this paper, the training dataset is used in building the classical OLS, regularized Ridge, Lasso and Elastic net models.

Estimating the Models Using Breast Cancer Data

The Breast Cancer Data

In this study, observations of breast cancer data are used and were sourced from Ahmadu Bello University, Teaching Hospital (ABUTH) Zaria, with the following exposure (or predictor) variables. Apart from the intercept, the other variables used in the analysis are age (Age20-29, Age30-39, Age40-49, Age50-59, Age60-69, Age70+), sex (male, female), marital status (married, single) and stage of the disease (stage1, stage2, stage3). For the purpose of analysis, all the 13 predictor variables are coded 1 for the presence of the event and zero otherwise, except for the dependent variable where its natural log is used. The intercept is coded 1. The study duration was for 60 months after having been diagnosed with breast cancer and the survival times in months of the patients is considered as the response variable. The object of the cancer study is to find out which of the predictor variables exert greater influence on the survival of the breast cancer patients, and using the exposure variables to predict the breast cancer patient's survival time. The sample collected consist of 312 breast cancer patients, and the study subject included 299 females (95.8%) and 13 males (4.2%) with an average age of 43.1 (with standard deviation of 11.7) for females and average age of 48.5 (with standard deviation of 12.0) for males. All the ages range between 20 and 75 years. The 5-year average survival time in months for females and males after being diagnosed with breast cancer were found to be 52.5 and 38.5 respectively. The data is divided into two subsets. The first subset is the training set (n_{train} =200 subjects) that will be used in estimating the models, while the second subset is the testing set (n_{test}= 112 subjects), which will be used for assessing the prediction accuracy of the estimated models.

Estimating the Regression Models

The regression models estimated parameters were obtained from the training data. Using all the predictors and the intercept, the p x p matrix X^TX is singular and so does not have an inverse. As a result, we cannot estimate β of equation (4) and the OLS model. The Ridge regularised estimate of β of equation (7), with the determined optimal value of λ_{opt} = 0.01 as shown in Fig 4(a), its standard error, the t-values and the p -values are presented in Table 1(a).

Table 1(a) : Ridge Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	0.9024	0.3387	2.6644	0.0084
Age20-29	0.0701	0.1987	0.3529	0.7245
Age30-39	0.1789	0.1430	1.2511	0.2125
Age40-49	0.1432	0.1413	1.0131	0.3123
Age50-59	-0.0336	0.1659	-0.2023	0.8399
Age60-69	0.1316	0.2141	0.6146	0.5396
Age70-79	0.4121	0.3134	1.3150	0.1901
Male	0.6327	0.6712	0.9426	0.3471
Female	0.4357	0.6673	0.6530	0.5146
Single	0.3761	0.1769	2.1263	0.0348
Married	0.5263	0.1928	2.7291	0.0070
Stage 1	1.3566	0.2388	5.6808	0.0000
Stage 2	0.2614	0.1592	1.6414	0.1024
Stage 3	-0.7156	0.1549	-4.6202	0.0000
Mean Squared Error				0.8239
R Squared				0.3227
Lamda opt				0.0100
Residual Standard error				0.9412

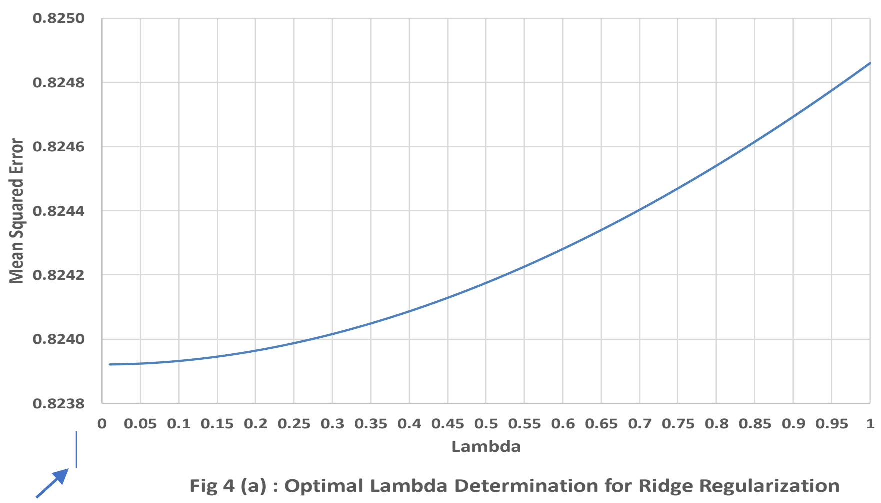

Fig 4 (a) : Optimal Lambda Determination for Ridge Regularization

From Fig. 4(a) the different values of lambda (λ) are on the x-axis, the least value of lambda (λ) is the optimal and is obtained at the beginning of the curve shown by the arrow i.e between 0.0 to 0.05, and λ_{opt} = 0.01

The Lasso regularized estimate of β of equation (16), with the determined optimal value of λ_{opt} = 0.04 as shown in Fig 4(b), its standard error, the t-values and the p -values are presented in Table 1(b).

Table 1 (b): Lasso Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	0.2406	0.0088	27.3912	0.0000
Age20-29	-0.0605	0.0571	-1.0591	0.2909
Age30-39	0.0538	0.0631	0.8514	0.3957
Age40-49	0.0095	0.0003	32.7941	0.0000
Age50-59	-0.1646	0.2085	-0.7891	0.4311
Age60-69	-0.0187	0.0012	-15.0659	0.0000
Age70-79	0.2783	0.4195	0.6632	0.5080
Male	0.6340	0.7643	0.8296	0.4079
Female	0.4369	0.7635	0.5722	0.5679
Single	1.5169	0.7801	1.9446	0.0533
Married	1.6669	0.7964	2.0932	0.0377
Stage 1	1.0066	0.3014	3.3397	0.0010
Stage 2	-0.0872	0.0004	-218.5425	0.0000
Stage 3	-1.0638	0.1411	-7.5380	0.0000
Mean Squared Error				0.8240
R Squared				0.3226
Lamda opt				0.0400
Residual Standard error				0.9413

Table 1 (c): Elastic net Regularised Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	0.3207	0.0142	22.5253	0.0000
Age20-29	0.2450	0.2236	1.0955	0.2747
Age30-39	0.3538	0.1695	2.0874	0.0382
Age40-49	0.3181	0.1649	1.9288	0.0553
Age50-59	0.1420	0.0603	2.3543	0.0196
Age60-69	0.3063	0.2500	1.2250	0.2221
Age70-79	0.5864	0.3763	1.5583	0.1209
Male	0.6342	0.7197	0.8811	0.3794
Female	0.4373	0.7172	0.6097	0.5428
Single	1.1155	0.7466	1.4942	0.1368
Married	1.2657	0.7637	1.6574	0.0991
Stage 1	1.0226	0.3025	3.3807	0.0009
Stage 2	-0.0731	0.0002	-475.6012	0.0000
Stage 3	-1.0499	0.1414	-7.4229	0.0000
Mean Squared Error				0.8239
R Squared				0.3227
Lamda opt, alpha opt			0.02	0.0100
Residual Standard error				0.9412

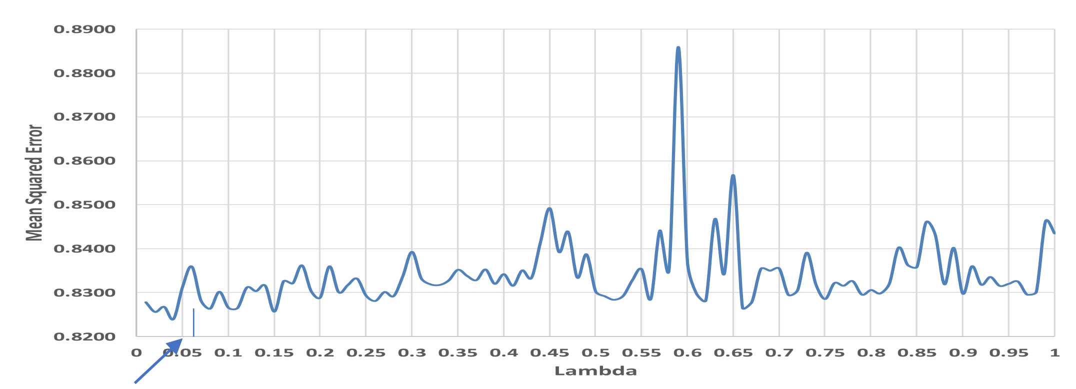

Fig 4 (b) : Optimal Lambda Determination for Lasso Regularization

From Fig.4(b) the different values of lambda (λ) are on the x-axis, the least value of lambda (λ) is the optimal and is obtained at the lowest part of the curve shown by the arrow i.e. between 0.0 to 0.05, and λ_{opt} = 0.04

The elastic net regularized estimate of β of equation (24), with the determined optimal value of λ_{opt} = 0.02 and α_{opt} =0.01 as shown in Figs 4(c) and 4(d), its standard error, the t-values and the p -values are presented in Table 1(c).

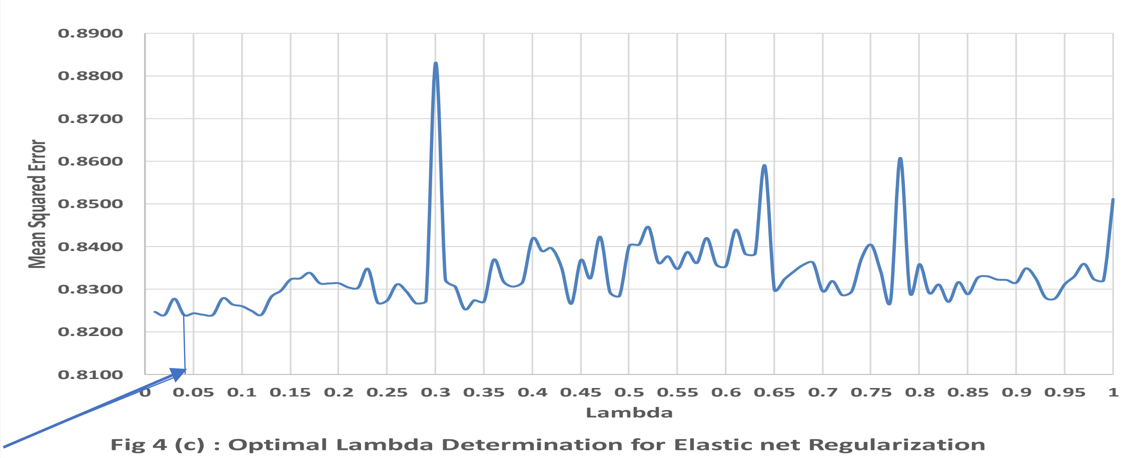

Fig 4 (c) : Optimal Lambda Determination for Elastic net Regularization

From Fig.4(c) the different values of lambda (λ) are on the x-axis, the least value of lambda (λ) is the optimal and is obtained at the lowest part of the curve shown by the arrow i.e between 0.0 to 0.05, and λ_{opt} = 0.02

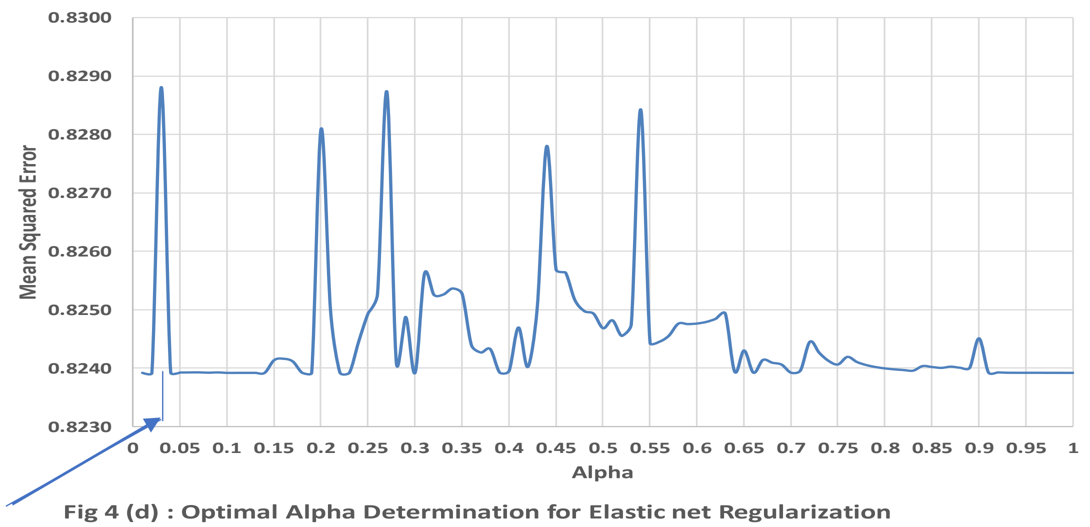

Fig 4 (d) : Optimal Alpha Determination for Elastic net Regularization

From Fig.4(d) the different values of alpha (α) are on the x-axis, the least value of alpha (α) is the optimal and is obtained at the beginning of the curve shown by the arrow i.e between 0.0 to 0.05, and α_{opt} = 0.01

While the Ridge regression results presented in Table 1(a) indicate five variables as statistically significant determinants of survival time of cancer patients, the Lasso results show seven significant variables. The elastic net model indicates that eight variables are significant in determining the survival times of the breast cancer patients. The results indicate that Age60-69, stages 2 and 3 of the disease tend to shorten the survival times of the patients. In contrast, Age30-39, Age40-49, Age50-59, married and stage 1 of the disease tend to increase the patient survival time. Based on the R^2 value, the Lasso model appears somewhat superior to the other two models.

When four of the variables namely: Age20-29, Male, Single and Stage 2 of the disease, are dropped from the analysis, the resulting matrix X^TX was able to have an inverse, and the resulting OLS estimated is shown in Table 2(a). Also, the regularized estimates for Ridge, Lasso and Elastic net are also given in Tables 2(b), 2(c) and 2(d), respectively. Apart from the lack of information on the other variables that were dropped from the analysis, all the results are very similar, indicating that the intercept, Stage 1 and Stage 3 of the disease are the only statistically significant variables in the determination of survival time of the breast cancer patients.

Table 2 (a): OLS Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	2.3842	0.2603	9.1590	0.0000
Age30-39	0.1163	0.2415	0.4814	0.6308
Age40-49	0.0809	0.2399	0.3372	0.7364
Age50-59	-0.0964	0.2608	-0.3696	0.7121
Age60-69	0.0694	0.3076	0.2258	0.8216
Age70-79	0.3460	0.4095	0.8449	0.3993
Female	-0.1736	0.1672	-1.0380	0.3006
Married	-0.1659	0.1470	-1.1283	0.2606
Stage 1	1.0876	0.2998	3.6282	0.0004
Stage 3	-0.9896	0.1395	-7.0933	0.0000
Mean Squared Error				0.8278
R Squared				0.3195
Residual Standard error				0.9335

Table 2 (b): Ridge Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	2.3826	0.2597	9.1751	0.0000
Age30-39	0.1171	0.2409	0.4863	0.6273
Age40-49	0.0817	0.2393	0.3413	0.7332
Age50-59	-0.0955	0.2602	-0.3671	0.7139
Age60-69	0.0703	0.3069	0.2292	0.8190
Age70-79	0.3463	0.4084	0.8477	0.3976
Female	-0.1729	0.1671	-1.0345	0.3022
Married	-0.1657	0.1470	-1.1272	0.2611
Stage 1	1.0868	0.2994	3.6292	0.0004
Stage 3	-0.9894	0.1395	-7.0940	0.0000
Mean Squared Error				0.8278
R Squared				0.3195
Lamda opt				0.0100
Residual Standard error				0.9335

Table 2 (c): Lasso Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	2.3915	0.2882	8.2991	0.0000
Age30-39	0.1087	0.3137	0.3466	0.7293
Age40-49	0.0736	0.3791	0.1941	0.8463
Age50-59	-0.1041	0.3592	-0.2898	0.7723
Age60-69	0.0594	0.2206	0.2693	0.7880
Age70-79	0.3378	0.4477	0.7545	0.4515
Female	-0.1752	0.1712	-1.0232	0.3075
Married	-0.1647	0.1536	-1.0721	0.2850
Stage 1	1.0871	0.3017	3.6033	0.0004
Stage 3	-0.9892	0.1407	-7.0328	0.0000
Mean Squared Error				0.8278
R Squared				0.3195
Lamda opt				0.0100
Residual Standard error				0.9335

Table 2 (d): Elastic net Regression of Breast Cancer Data

Variable	Estimate	SE	t-value	P-value
Intercept	2.3835	0.3439	6.9310	0.0000
Age30-39	0.1166	0.4094	0.2848	0.7761
Age40-49	0.0812	0.5122	0.1585	0.8742
Age50-59	-0.0961	0.5441	-0.1765	0.8601
Age60-69	0.0698	0.3982	0.1753	0.8610
Age70-79	0.3461	0.5090	0.6799	0.4974
Female	-0.1733	0.1747	-0.9920	0.3225
Married	-0.1657	0.1607	-1.0309	0.3039
Stage 1	1.0873	0.3026	3.5935	0.0004
Stage 3	-0.9894	0.1424	-6.9490	0.0000
Mean Squared Error				0.8278
R Squared				0.3195
Alpha opt, Lambda opt			0.19	0.0200
Residual Standard error				0.9335

The determined optimal λ_{opt} = 0.02 and α_{opt} =0.19 as shown in Figs 5(a) and 5(b), for the Elastic net regularization that resulted in Table 2(d).

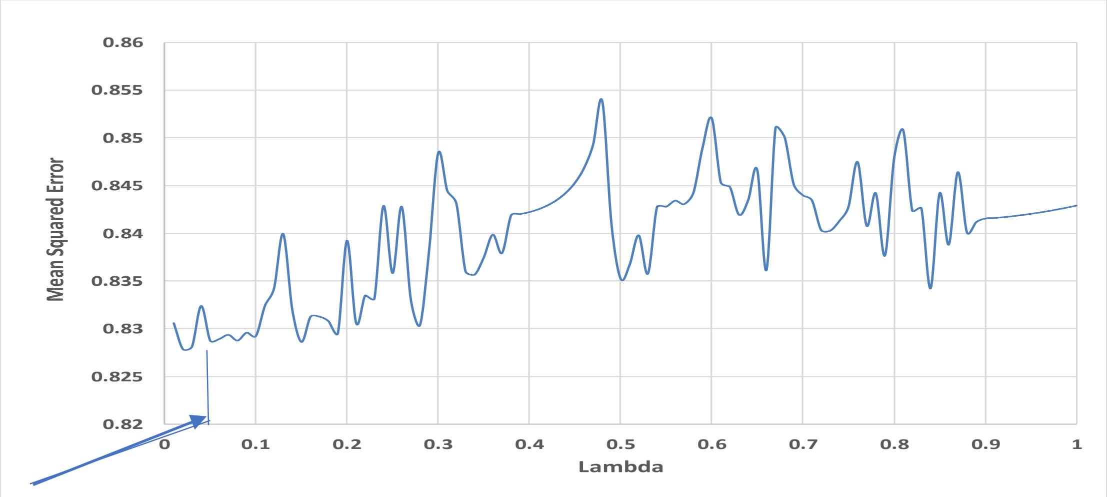

Fig 5(a): Optimal Lambda Determination for Elastic Net Regularization

From Fig.5(a) the different values of lambda (λ) are on the x-axis, the least value of lambda (λ) is the optimal and is obtained at the lowest part of the curve shown by the arrow i.e between 0.0 to 0.1, and λ_{opt} = 0.02

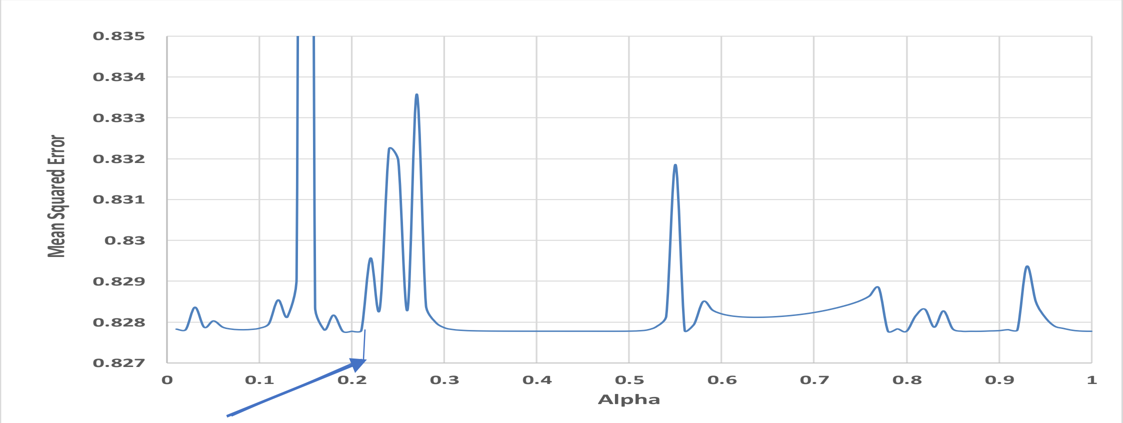

Fig 5(b): Optimal Alpha Determination for Elastic Net Regularization

From Fig.5(b) the different values of alpha (α) are on the x-axis, the least value of alpha (α) is the optimal and is obtained at the lowest part of the curve shown by the arrow i.e between 0.1 to 0.2, and α_{opt} = 0.19

Prediction Accuracy of the Estimated Models

We shall apply the estimated models based on the training dataset to assess the prediction accuracy of the regularized Ridge, Lasso and Elastic net as compared to the OLS model using the testing set comprising of n_{test}= 112 patients. As we have seen in Section 3, the OLS model could not be estimated if all the variables are included because of multi-collinearity. This suggests that when all the variables are included, the regularized Ridge, Lasso or Elastic net should be used. Based on the estimated parameters presented in Tables 1(a), 1(b) and 1(c), we predict the log of the survival times of the 112 patients that are used as testing set using the models:

$$\widehat{Y}_{(\lambda,\alpha)} = X\widehat{\beta}_{(regularised)} \qquad (28)$$

and the associated MSE values

$$MSE(\lambda,\alpha) = (Y-\widehat{Y}_{(\lambda,\alpha)})^T(Y-\widehat{Y}_{(\lambda,\alpha)})/\,\mathrm{n}_{\mathrm{test}} \qquad (29)$$

The regularized model can be either Lasso or Ridge when α = 0 and λ > 0 and Elastic net when both λ and α are greater than zero. The R version 4.1.1 package was used for all the computations and the developed R codes are with the authors and available on request. The predicted results for Ridge, Lasso and Elastic net when all the variables and the constant are included are presented in Figs 6(a), 6(b) and 6(c), respectively.

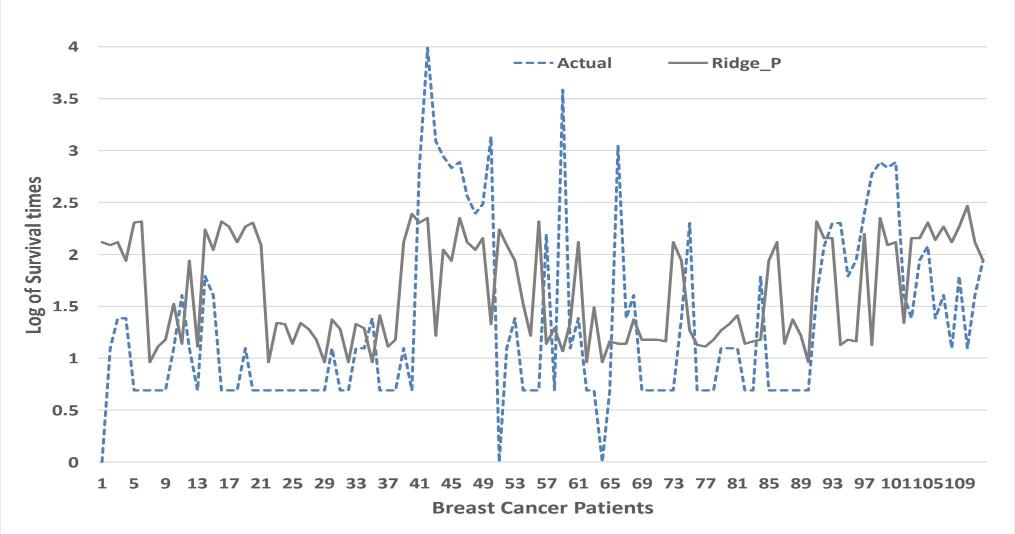

Fig 6 (a): Predicted and actual log of Survival times using Ridge Regularization

Fig. 6(a) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using Ridge regression model.

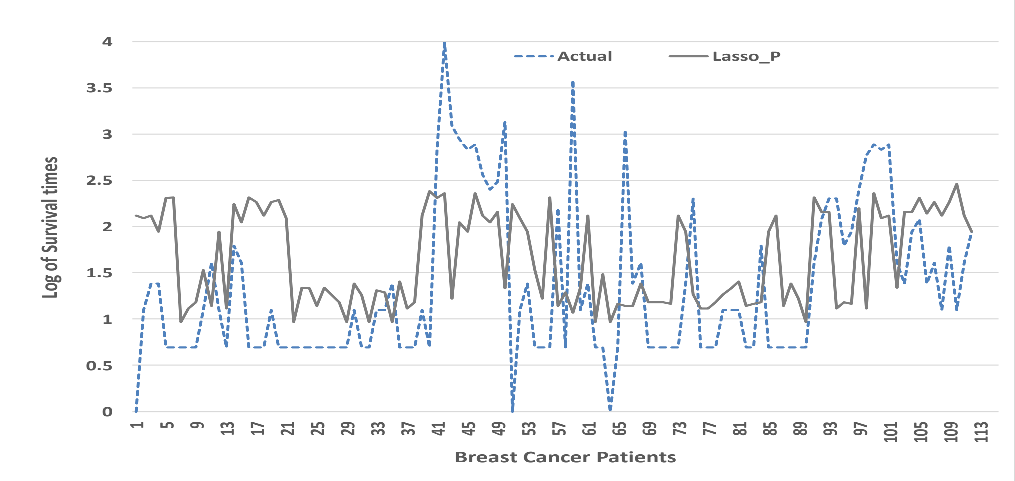

Fig 6 (b): Predicted and actual log of Survival times using Lasso Regularization

Fig. 6(b) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using Lasso regression model.

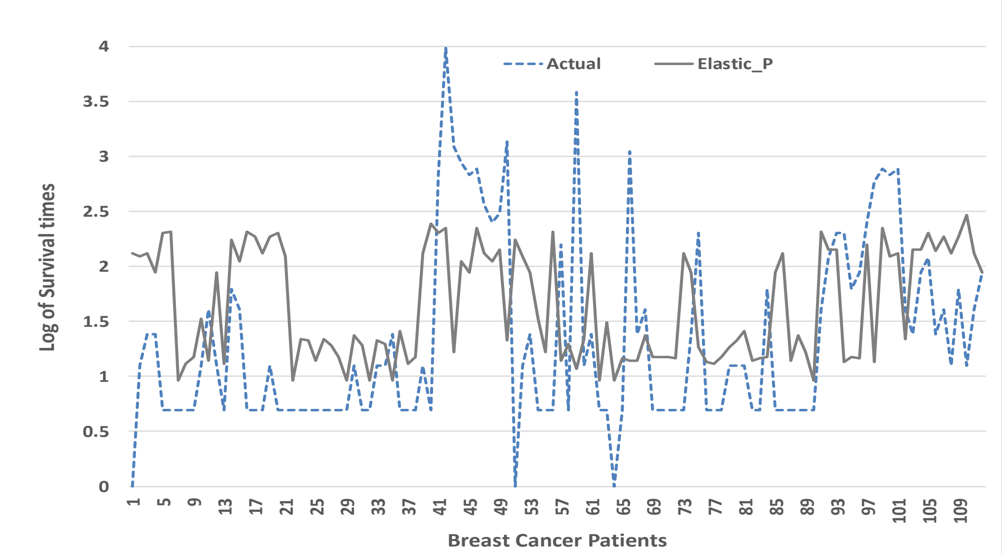

Fig 6 (c): Predicted and actual log of Survival times using Elastice net Regularization

Fig. 6(c) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using Elastic net regression model.

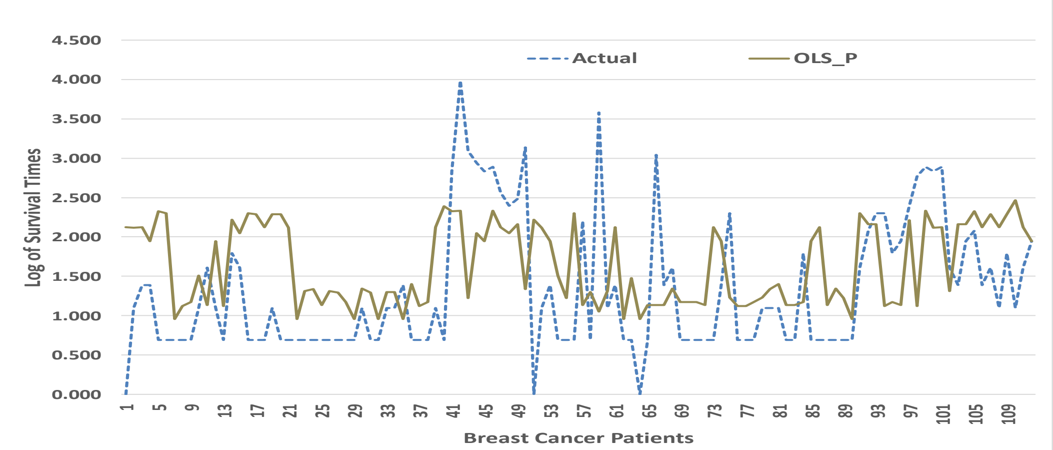

Fig 7(a): Actual and Predicted Log of Survival Times Using OLS Regression

Fig. 7(a) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using OLS regression model after dropping four variables (Age20-29, Male, Single and Stage2).

Using the computed MSE values, the Lasso model produced the least MSE of 0.832178, followed by Elastic net model with MSE value of 0.83274, with the Ridge model reporting the highest MSE value of 0.83283. From these results it is clear that the Lasso model would be preferable to the other two models. However, with the four variables: - Age20-29, Male, Single and Stage2, of the disease dropped, the predicted and the actual values are presented in Figs 7(a), 7(b), 7(c) and 7(d), respectively for the OLS, Ridge, Lasso and Elastic net models.

The Lasso model maintains its superiority over the other three models with computed MSE value of 0.833866. This is followed by Ridge with MSE value of 0.834409, Elastic net with MSE value of 0.834471 and the OLS model with MSE value of 0.834507. Thus, using the reduced model, the three regularized models are all superior to the OLS model.

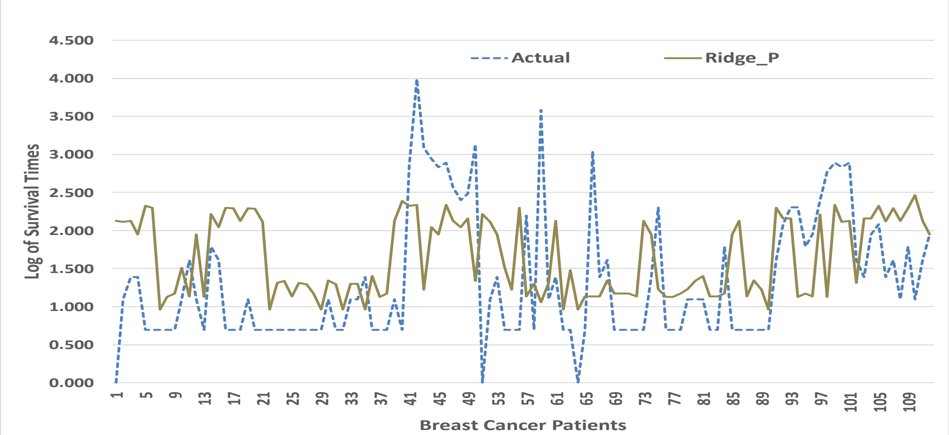

Fig 7(b): Actual and Predicted Log of Survival Times Using Ridge Regression

Fig. 7(b) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using Ridge regression model after dropping four variables (Age20-29, Male, Single and Stage2).

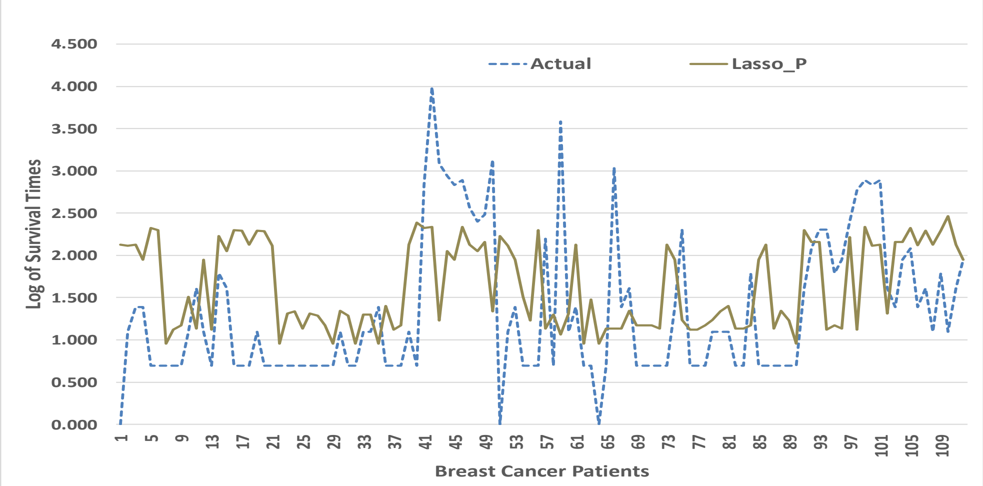

Fig 7(c): Actual and Predicted Log of Survival Times Using Lasso Regression

Fig. 7(c) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using Lasso regression model after dropping four variables (Age20-29, Male, Single and Stage2).

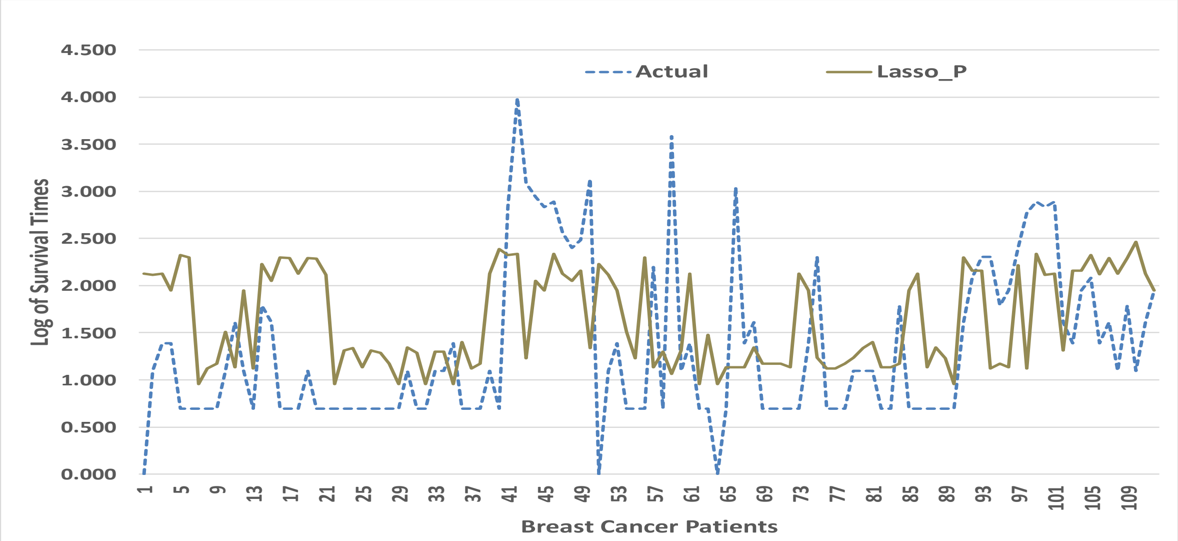

Fig 7(d): Actual and Predicted Log of Survival Times Using Elastic net Regression

Fig. 7(d) depict the plot of log of survival times against the breast cancer cases for actual and predicted values using Elastic net regression model after dropping four variables (Age20-29, Male, Single and Stage2).

CONCLUSION

The Ridge, Lasso and Elastic net regularized models outperform the linear regression model in terms of prediction accuracy and information content on the predictor variables. The regularised models by design ensured that all the predictor variables can be used in estimating the model. On the regularized models, Lasso model appeared superior to Ridge and Elastic net models. The results of our study have shown that of the 14 variables used, eight are significant factors of breast cancer determination. In this study, we found that breast cancer patients in age group 60-69, that are in Stage 2 and Stage 3 of the disease have lower survival times and therefore have higher risk of dying from the disease. In contrast, patients that are either single or married and are in Stage 1 of the disease have longer survival times and hence lower risk of dying from the disease.

It is, therefore, recommended that the Federal and State Ministries of health should embark and sustain awareness campaigns about the breast cancer in the population in order for the sufferers to be detected and treated early, so as to improve the survival status.

REFERENCES

Ahrens A., Schaffer M. E. and Hansen C. B. (2018). Prediction, Model Selection and Causal Inference with Regularised Regression. https://statalasso.github.io/

Bray F. J., Ferley J., Soerjomataram I., Siegel L, Torre A. and Jamal A. (2018). Global Cancer Statistics. Cancer Journal of Clinicians vol. 68, no.6 pp394-424

Chaturvedi N. (2018). Statistical Modelling for Integrative Analysis of Multi-Omics DataVU University Medical Center-Cancer Center Amsterdam. PhD Thesis. Unpublished

Emmert-Streib F. and Dehmer M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularisation, Shrinkage, (http:// creativecommon.org/by/4.0/).

Fitzmaurice C. (2018). A systematic Analysis for the Global Burden of Disease Study. *JAMA Oncology vol*. 36

Frank E. and Matthias D. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection.

Goeman, J. J. (2010). L1 penalized estimation in cox proportional hazards model*. Biometrical Journal,* 52(1):70 { 84. doi 10.1002/bimj200900028.

Guyon and Elisseeff, (2003). An Introduction to Variable and Feature Selection *Journal of Machine Learning Research* 3: 1157-1182.

Ho, M.; Sun, Z.; and Xin, J. (2015). Weighted elastic net penalized mean-variance portfolio design and computation. SIAM J. Finance. Math., 6, 1220–1244

Hoerl A. E. and Kennard R. W. (1970). Ridge regression: Bias estimation for nonorthogonal Problem*, Technometrics,* 12, 55-67

Liu, W. and Li, Q. (2017). An efficient elastic net with regression coefficients method for variable selection of spectrum data. DOI:10.1371/journal.pone.0171122

Liv R., Xiaocen L. and Lam K. S. (2017). Combinatorial Chemistry in Drug Discovery. HHS Public Access. *Doi:*10.1016/j.cbpa.2017.03.017

Madjar K. (2018). Survival models with selection of genomic covariates in heterogeneous cancer studies. PhD Dissertation, Unpublished

Melkumova L. E. and Shatskikh S. Y. (2017). Comparing Ridge and Lasso Estimators For Data Analysis. 3rd International Conference "Information and Technology And Nanotechnology, Samara Russia. *ELSEVIER*

Park, H. and Konishi, S. (2015). Robust logistic regression modelling via the elastic net-Type regularization and tuning parameter selection *Journal of Statistical Computation and Simulation,* 86(7): 1-12.

Pavlou M., Ambler G. Seaman S., De Iorio M. and Omar R. Z. (2015). Review and Evaluation of Penalised Regression Methods for Risk Prediction in Low-Dimentional Data with Few Events. *Statistics in Medicine*. Wiley Online Library.

PSU, P. S. U. (n.d). Ridge regression. Retrieved from https://onlinecourses.science.psu.edu/stat857/ node/155/

Rebecca, A. H., Jonathon, J. V., Sarah, C. E., Matthew, G. B., and Thomas, G. D. (2015). Penalized likelihood methods that improve parameter estimates in occupancy models. 6: 949-959. dois: 10111/2041-210x.12368.

Shen G., Kang T., Yang S., Baek S., Jeong Y. and Kim S. (2011). GENT: Gene Expression Database of Normal and Tumor Tissues. *Cancer Informatics*.

Statistical Learning-Lasso. Retrieved from https://datacadamia.com/data_mining/lasso

Steele, A.J.; Cakiroglu, S.A.; Shah, A.D.; Denaxas, S.C.; Hemingway, H.; Luscombe, N.M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease.

Talib R., Hanif M. K., Ayesha S., and Fatima F. (2016). Text Mining: Techniques, Applications and Issues. *Inter. Journal of Advanced Computer Science & Applis*, Vol. 7 No.11.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B (methodological), 58(1):267-288.

Tuji, M. A. (2010). Variable Selection in Cox-Models using the L1-Regularization Path Algorithm. Norwegian University of Science and technology.

Xiao, N., Xu, Q., and Li, M. (2016). Hdmon: building nomograms for penalized cox models with high-diemnsional survival data Doi:http://dx,doi.org/10,1101/065524.

Van der Kooij, A.J. and Meulman, J. .J.(2006). Regularization with Ridge penalties, the Lasso, and the Elastic Net for Regression with Optimal Scaling Transformations.

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalizedregression. *Journal of Mathematical Psychology*, 89, pp 1–10.

Wessel N. and Van Wieringen (2020). Lecture Notes on Ridge Regression. http://creativecommons.org/licenses/by-nc-sa/4.0/ pp 5-12

Yichao W. and Yufeng L. (2009). Variable Selection in Quantile Regression. *Statistica Sinica* Vol. 19 . Institute of Statistical science pp 801 – 817

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society B*, 67(2):301-320.