BENIN JOURNAL OF
STATISTICS
ISSN 2682-5767
Vol. 5, pp. 61– 74 (2022)

Regularized Regression Model for Predicting Hypertension and Type 2 Diabetes Mellitus in Patients

D. M. Okewole[1], **R. J. Salako**[2], **K. S. Adekeye**[3] * **and S. O. Adesanya**[4]

[1,3,4]*Department of Mathematical Sciences, Redeemer's University, Ede, Nigeria*

[2]*Department of Statistics, Osun State Polytechnic, Iree, Nigeria.*

(*Received: 09 July 2021; Accepted: 22 August 2021*)

Abstract. Regularized Regression methods such as Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression are some of the techniques that overcome ordinary least squares assumptions' violation such as multicollinearity. Modeling hypertension and diabetes involve several explanatory variables, some of which are interrelated. Thus, there is the need to use an estimation technique that can solve the problem of interrelated variables in modeling. The consideration of hypertension and diabetes in this study is on the premise that the two are related and have some predictors in common. There were four dependent variables in the study: Fasting Blood Sugar (FBS), urea, Systolic Blood pressure (SBP) and Diastolic Blood Pressure (DBP) and thirteen independent variables. Comparisons were made using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The results showed that the model on SBP had the best performance in the final LASSO models which retained Height, Religion, Age, Sex, Marital Status, Creatinine, Family History, Temperature and Type of disease, as well as the ridge regression model. One of the implications of the result is that certain levels of these independent variables can imply the levels of the dependent variables that signify the presence of type 2 diabetes or hypertension. The LASSO method performed better than Ridge regression for FBS, urea and SBP. With both LASSO and ridge regression, multicollinearity problem in the independent variables was removed.

Keywords: LASSO, Hypertension, Diabetes, Tuning, Shrinkage penalty, Cross validation

Published by: Department of Statistics, University of Benin, Nigeria

1. Introduction

Regularized Regression methods such as Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression, are some of the techniques that

*Corresponding author. Email: adekeyek@run.edu.ng

overcome ordinary least squares assumptions' violation such as multicollinearity. Modeling hypertension and diabetes involve several explanatory variables, some of which are interrelated, thus requiring an estimation technique that can deal with such situations. The consideration of hypertension and diabetes in this study is on the premise that the two are related, having some predictors in common.

Hypertension is known to be a blood pressure more than 140mm Hg and 90mm Hg for systolic and diastolic pressures, respectively (WHO, 2019), while Type 2 diabetes mellitus is high glucose blood content outside the normal range of 70 - 110 mg/dl, following an overnight 8 to 10 hours fast (Davidson, 2001). Both hypertension and diabetes are very deadly diseases and can develop further from the pre-stages if not properly managed. The contribution of non-clinical diagnosis in identifying significant predictors could be an important step in managing these diseases.

The global trend of hypertension and diabetes has been deteriorating despite the wealth of knowledge available on their prevention and management. In 2010, the estimated global prevalence of hypertension was more than 25% among adults, and it is projected to rise, especially in developing countries like Nigeria. Kearney *et al.* (2005) and Chataut *et al.* (2011), reported that there is an epidemiological shift in the prevalence of hypertension in developing countries as compared to developed countries. Cappuccio and Miller (2016) submitted that by the year 2025, the proportion of world's adult population that will be affected with hypertension is likely to reach 29% of world population.

Research involving modeling hypertension as well as diabetes abounds in literature. Saebom *et al.* (2019) considered hypertension and Type 2 diabetes using four procedures which led to the production of quantitative maps with paths linking the Single-Nucleotide Polymorphisms (SNPs) with Hypertension through phenotype and type 2 diabetes. Ramezankhani *et al.* (2019) used Cox regression to investigate associations of marital status with Type 2 diabetes, hypertension, and cardiovascular disease. Tuoyire and Ayeteh (2018) used binary logistic regression to study the association of marital status with Type 2 diabetes and hypertension. A study on marital differences in blood pressure and risk of hypertension among Polish men using multifactorial models of logistic regression by Anna and Monika (2005) showed that systolic blood pressure and diastolic blood pressure in never-married men is higher than married men.

Hilawe *et al.* (2013), in the investigation of differences by sex in the prevalence of diabetes mellitus, concluded that women in sub-Saharan African countries are more likely to be obese and consequently have a greater prevalence of diabetes than men. Chataut *et al.* (2011) identified blood glucose level, Age, Gender, Literacy, Physical inactivity, Body mass index (BMI), smoking, and alcohol consumption as factors that contributed significantly to casualties in people with diabetes. Cardiovascular risk among hypertensive and normotensive subjects having Type 2 diabetes mellitus by patients attending medical checkups between 1992 and 2011 was studied by Michel *et al.* (2017) using Cox regression. Amanda (2020) compared LASSO, ridged, and elastic regressions in predicting Type 2 diabetes. Hu and Zhang (2020) isolated influencing factors in diabetic nephropathy in obese patients with Type 2 Diabetes Mellitus using logistic regression in LASSO.

Most of the research on modeling in the literature does not include the case of specifying quantitative dependent variables on hypertension and diabetes for the purpose of obtaining significant predictors. This was considered in this study. The essence is to be able to present the identified predictors for further studies in improving the prediction performance of the hypertension and diabetes measurement variables. The important predictors were identified using LASSO regression which shrinks irrelevant coefficients to zero and compared with results from Ridge regression which presents all coefficients in the final model. There were four dependent variables on hypertension and Type 2 diabetes mellitus and thirteen independent variables using data collected from two hundred and thirty-four (234) hypertension and diabetes patients at clinical levels. The dependent variables are: Fasting Blood Sugar, Urea, Systolic Blood Pressure, and Diastolic Blood Pressure while the independent variables are: weight, height, religion, age, sex, marital status, creatinine content, dizziness, headache, family history, temperature, type, and Body Mass Index.

2. Materials and Method

2.1 *Least Absolute Shrinkage and Selection Operator*

Least Absolute Shrinkage and Selection Operator (LASSO) selection arises from a constrained form of ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter. LASSO estimates the regression coefficients by maximizing the log-likelihood function with the constraint that the sum of the absolute values of the regression coefficients is less than or equal to a positive constant (Ateeq *et al.,* 2019). The main aim of the LASSO technique is to reduce the regression coefficients as much as possible by setting some regression coefficients exactly to zero (Milani *et al.,* 2016). As an extension to classical or traditional regression, the LASSO regression coefficients are constrained by the shrinkage penalty.

The LASSO regression performs both variable selection and regularization in other to enhance the prediction and interpretability of the statistical regression model.

The LASSO estimate of regression solution is given by

$$arg\min_{\beta}\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{n}\beta_j x_{ij}\right)^2\} \tag{1}$$

subject to

$$\|\beta\|_1 \leq t \tag{2}$$

(2) is the constraint for the regression model, and t is a tuning parameter (also called regularization parameter or penalty term), and $\|\beta\|_1$ is the L_1-norm.

(1) in the Lagrange form is presented as

$$arg\min_{\beta}\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{n}\beta_j x_{ij}\right)^2\}+\lambda\|\beta\|_1 \tag{3}$$

It should be noted that the LASSO regression adds a factor of sum of absolute value of coefficients in the optimization objective. In (3), the $RSS = \sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{n}\beta_j x_{ij}\right)^2$ and $\lambda\|\beta\|_1$ is the shrinkage penalty. The LASSO estimate of regression solution in (3) was rewritten by Emmert-Streib and Dehmer (2019) as

$$\hat{\beta} = arg\min_{\beta}\{\frac{1}{2n}\|y-X\beta\|_2^2+\lambda\|\beta\|_1\} \tag{4}$$

It should be noted that λ in (4) can take various values which implies different interpretations. When $\lambda = 0$, then the shrinkage penalty has no effect, and LASSO regression model produces the same coefficients as OLS. However, as λ approaches ∞, the shrinkage penalty becomes more influential, and the predictor variables that are not importable in the model get shrunk towards zero, and some even get dropped from the model.
The sparsity of the coefficients of the regression increases as λ approaches ∞. Thus, the computational and theoretical results are difficult to obtain in LASSO because it lacks analytical solution. To solve the problem of sparsity, the use of the machine learning approach is employed in this work.

2.2 *Ridge Regression*

The objective function of the Ridge regression is similar to that of LASSO but the constraints are different. The constraints for the Ridge regression model are given as:

$$arg\min_{\beta}\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{n}\beta_j x_{ij}\right)^2\} \tag{5}$$

subject to

$$\|\beta\|_2^2 \leq t \tag{6}$$

where t is a tuning parameter (also called regularization parameter or penalty term) and $\|\beta\|_2^2$ is the l_2-norm.
(5) can be written in the Lagrange form given as:

$$arg\min_{\beta}\{\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\} \tag{7}$$

$$\hat{\beta} = arg\min_{\beta}\{\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\} \tag{8}$$

2.3 *Prediction Performance Metrics*

Performance metrics (error measurement) is an important aspect of modelling. In machine learning regression, prediction performance metrics is a method used to compare the trained model predictions with the observed data from the testing data set. Various performance metrics are available in the literature, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-Squared (R^2), Adjusted R-Square ($\overline{R^2}$) , Mean Square Percentage Error (MSPE), Root Mean Squared Logarithmic Error (RMSPE), and many more. In this paper, the MSE and RMSE are used for measuring prediction performance of the models. The underlying assumption when presenting MSE and RMSE is that the errors are unbiased and follow a normal distribution (Chai and Draxler, 2014). Thus, Using the MSE and RMSE helps to provide a complete picture of the error distribution.

2.4 *K-Fold Cross-Validation*

Following standard literature procedures, optimal shrinkage parameters were tunes for the LASSO-based regularized as well as Ridge regression multi-response gaussian model. The R- software (R Core Team, 2019) was used to obtain the multi-response Gaussian family using family = *"mgaussian"* option in *glmnet* package.
The matrix of quantitative responses which are: FBS, Urea, Diastolic BP and Systolic BP were utilized in the regularized regression multi-task learning on the explanatory variable matrix X with the optimal values of tuning parameter $\hat{\lambda}_{min}$. The resulting multi-response Gaussian models serve as predictive models for hypertension profiles and type II diabetes profiles in patients, respectively.
In the model building step, we first partitioned the study data into training and test sets using the 80:20 scheme. The penalized multi-response Gaussian regression model was developed using the 80% (187 samples) training data and 20% (47 unseen samples) was the test data. Prediction performance of the developed model was evaluated using MSE as diagnostic metrics.

2.5 *Cross Validating*

In high dimensional problems, Cross-Validation (CV) is used to select the regularization parameter to determine the best regression model in a study. CV is a model evaluation technique which prevents over-fitting by applying the model

to the data that are not involved in the fitting.
In the LASSO regression step, λ is a tuning parameter that needs to be estimated via cross-validation. The means-squares error (MSE) for each CV fold, say F_k, is estimated by

$$e(\lambda)_k = \frac{1}{\#F_k} \sum_{j \in F_k} (y_i - \hat{y}_j)^2 \tag{9}$$

Where, $\#F_k$ is the number of samples in set F_k. The average over all K folds is taken.

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} e(\lambda)^k \tag{10}$$

The expression in (10) is called cross-validation of the Mean-Squared Error (MSE).
To obtain an optimal λ from $CV(\lambda)$, two approaches are common. The first estimates the λ that minimizes the function $CV(\lambda)$.

$$\hat{\lambda}_{min} = argmin CV(\lambda) \tag{11}$$

The second approach begins with an estimation of $\hat{\lambda}_{min}$ and then identifies the maximal λ that has a cross-validation MSE smaller than $CV(\hat{\lambda}_{min} + SE(\hat{\lambda}_{min}))$. Thus,

$$\hat{\lambda}_{1se} = max \lambda_{CV(\lambda) \leq CV(\hat{\lambda}_{min}) + SE(\hat{\lambda}_{min})} \tag{12}$$

3. Results and Discussion

The data used in this study involved a group of patients with one or a combination of hypertension and Type 2 diabetes mellitus at clinical levels. The data is cross-sectional, and it was collected on 234 diabetes and hypertensive patients through the assistance of specialist doctors in the General Outpatient Department (GOPD) and as well as the record section of the college of medicine of Ladoke Akintola University of Technology (LAUTECH), Osogbo. The coding of the categories is exactly as it is contained in the medical records at LAUTECH Teaching Hospital.
The hypertension data (Systolic and Diastolic pressures) were measured using a BP apparatus which consists of an arm cuff, dial, pump, and valve. The Blood Glucose content was measured after at least eight hours of fasting (Fasting blood sugar test) with Glucometer.
There are four response variable indicators including Y_{11}= Fasting blood sugar (FBS), Y_{12}=Urea on Type 2 diabetes mellitus T_2DM, Y_{21}= Systolic blood pressure, and Y_{22}=Diastolic blood pressure on hypertension status using weight

(X_1), height (X_2), religion affiliation (X_3), age (X_4), sex (X_5), marital status (X_6), creatinine (X_7), dizziness (X_8), headache (X_9), family history (X_{10}), temperature (X_{11}), BMI (X_{12}), and disease type (X_{13}) as explanatory variables. The categories of disease type (X_{13}) are; Group 1: Renal disease, hypertension and diabetes, Group 2: diabetes, Group 3: hypertension, Group 4: Hypertension and diabetes, Group 5: Renal disease, Group 6: Renal disease and diabetes, Group 7: Renal disease and Hypertension.
The details of the other nominal variables in the study are as follows:

Religion affiliation (X_3): 1- Christianity, 2- Islam
Sex (X_5): 1- Female, 2- Male
Marital status (X_6): 1- Married, 2 – Single
Dizziness (X_8): 1- No, 2- Yes
Headache (X_9): 1- No, 2- Yes
Family (X_{10}): 1- No, 2- Yes
The Boxplots in Figure 1 indicated presence of outliers in the data. Therefore, the data was cleaned by the capping method, whereby values above the upper 1.5^*IQR limit were replaced with the 95th percentile and those below the lower 1.5^*IQR limit by the 5th percentile before proceeding to model the data.

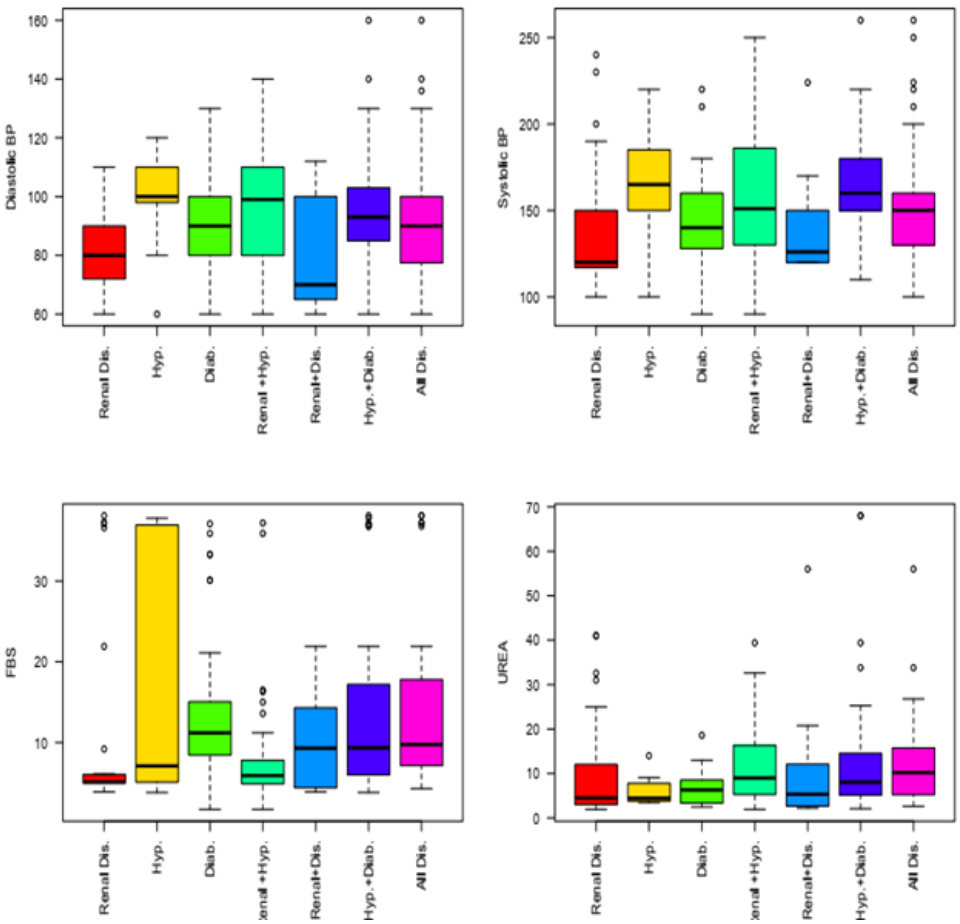

Figure 1: Boxplot Showing Response Variables versus Diagnosis Type in Patients.

The outlook of the data in its raw form as well as in the transformed form (by standardization), are contained in Tables A1 and A2 in the Appendix.

A check of the multicollinearity status of the data using Variance Inflation Factor (VIF) was carried out and the results are presented in Table 1. This is to show a major limitation of the least squares technique that is taken care of by regularized regression. A VIF near 1 suggests that multicollinearity is not a problem for the independent variables, whereas a VIF much greater than 1 indicates the presence of multicollinearity. It should be noted that a maximum VIF value more than 10 is often taken as an indication that the multicollinearity may be severe or unduly influencing the least square estimates.

Table 1: Variance of Inflation Test

2*Model		Collinearity Statistics	
		Tolerance	VIF
1	Weight	0.017	57.965**
	Height	0.102	9.786*
	Religion	0.955	1.047
	Age	0.443	2.259*
	Sex	0.843	1.186
	Marital	0.549	1.822
	Creatinine	0.806	1.241
	Dizziness	0.27	3.700*
	Headache	0.28	3.577*
	Family History	0.752	1.329
	Temperature	0.878	1.139
	Body Mass Index	0.023	43.567**
	Type	0.342	1.0621

Note: ** *indicates the variables with severe multicollinearity, and * indicates variables with multicollinearity.*

Figure 2 shows the LASSO $\hat{\lambda}_{min}$ parameter tuning path using the MSE criterion. The optimal value of $\hat{\lambda}_{min}$ is obtained at the vertical dashed lines where "log" $(\hat{\lambda}_{min}) = log(0.1258925) \approx -2.07233$. Results are shown in dependence on the regularization parameter $log(\lambda)$. The numbers on top of the figure give the number of non-zero regression coefficients which is 9 for the optimal value of $\hat{\lambda}_{min}$.

With optimal shrinkage parameter $\hat{\lambda}_{min} = 0.1258925$ obtained using the *cv.glmnet* option of the R software package *glmnet*, the model parameters were estimated from the training model. The estimated non-zero coefficients of the fitted model are presented in Table 2.

It is crystal clear from results in Table 2 that out of the thirteen independent variables included in the data, the fitted LASSO-based regularized multiresponse Gaussian model has successfully selected only nine predictors which are Height, Religion, Age, Sex, Marital Status, Creatinine, Family History, Temperature and Type of disease as the only relevant predictors for prediction of Type 2 diabetes and hypertension profiles of patients. All other variables have been shrunken to zero in the model fitting step by shrinkage and regularization procedure of the LASSO. Based on the results in Table 2, the regression models for each response variable are as follows:

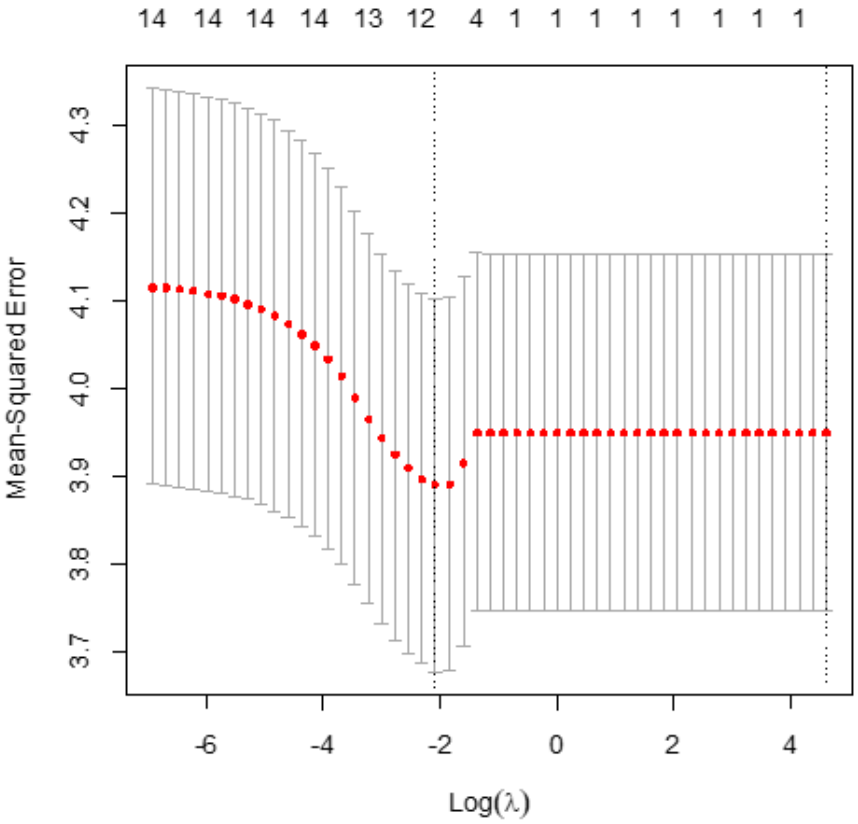

Figure 2: LASSO Shrinkage Parameter Tuning Path with MSE Criterion.

Table 2: LASSO Regularized Multi-response Regression Estimates

Parameter	FBS	Urea	Systolic	Diastolic
(Intercept)	0.4008	0.5061	0.0779	0.0596
Weight				
Height	0.0256	0.0303	0.0739	0.0665
Religion	-0.096	-0.2008	0.0167	0.0208
Age	0.0168	0.0211	0.0084	0.0119
Sex	-0.1015	-0.0495	0.0006	-0.0134
Marital_Status	-0.0217	-0.0706	-0.0645	-0.0568
Creatinine	0.0706	0.0934	0.0333	0.0231
Dizziness				
Headache				
Family History	-0.0013	-0.0046	0.0014	-0.0113
Temperature	-0.0005	-0.0011	0.008	0.0105
BMI				
Type	-0.0277	-0.0137	0.0017	-0.0071

$$\hat{FBS} = 0.4008 + 0.025X_2 - 0.0960X_3 + 0.0168X_4 - 0.1015X_5 - 0.02171X_6 + 0.0706X_7 - 0.0013X_{10} - 0.0005X_{11} - 0.0277X_{13} \quad (13)$$

$$\hat{Urea} = 0.5061 + 0.0303X_2 - 0.2008X_3 + 0.0211X_4 - 0.0495X_5 - 0.0706X_6 + 0.0934X_7 - 0.0046X_{10} - 0.0011X_{11} - 0.0137X_{13} \quad (14)$$

$$\hat{Sys} = 0.0779 + 0.0739X_2 + 0.0167X_3 + 0.0084X_4 + 0.0006X_5 - 0.0645X_6 + 0.0333X_7 + 0.0014X_{10} + 0.0080X_{11} + 0.0017X_{13} \quad (15)$$

$$\hat{Dias} = 0.0596 + 0.0665X_2 + 0.0208X_3 + 0.0119X_4 - 0.0134X_5 - 0.0568X_6 + 0.0231X_7 - 0.0113X_{10} + 0.0105X_{11} - 0.0071X_{13} \quad (16)$$

The Ridge regression on the other hand presents all the variables; it does not shrink any coefficient to zero. The results for the Ridge regression are presented in Table 3.

Table 3: Ridge Regression Estimates

Parameter	FBS	Urea	Systolic	Diastolic
(Intercept)	0.3362	0.4798	0.0618	0.2176
Weight	0.0068	0.0037	0.0265	0.0177
Height	0.013	0.0156	0.0409	0.0366
Religion	-0.057	-0.1189	0.0103	0.0125
Type	-0.0228	-0.0122	0.0013	-0.0062
Age	0.0249	0.0309	0.0112	0.0139
Sex	-0.0958	-0.0494	0.0008	-0.0157
Marital_Status	-0.0264	-0.0846	-0.0785	-0.0581
Creatinine	0.0391	0.0508	0.0168	0.0113
Dizziness	0.0037	0.0085	0.0035	-0.0132
Headache	-0.0093	-0.0441	0.016	-0.018
Family History	-0.0083	-0.0351	0.0136	-0.0964
Temperature	0.0003	-0.0018	0.025	0.0315
BMI	0.0159	0.014	0.0092	0.0224

To test for the prediction performances of the fitted models in (13) to (16), the multi-response regularized gaussian model was tested on the held-out test data set by plugging predictors in the 20% test (47 samples) to predict the corresponding response variables and then check means squared error (MSE) and the root mean -squared error (RMSE) of prediction, respectively. The prediction performance metrics for the fitted model are presented together with that of Ridge regression in Table 4.

From the values in Table 4, it is evident that the Systolic Blood Pressure model is best in terms of prediction performance of the LASSO method since it minimized both metrics the most. The next best model for LASSO is the urea, followed by the Diastolic Blood Pressure models, while the FBS model has the

Table 4: Prediction Evaluation Metrics

2*Metric	FBS		Urea		SYS		DIAS	
	LASSO	Ridge	LASSO	Ridge	LASSO	Ridge	LASSO	Ridge
MSE	1.106	1.1198	1.0353	1.0487	0.9917	1.0072	1.1035	1.0927
RMSE	1.0517	1.0582	1.0175	1.0241	0.9959	1.0036	1.0505	1.0453

highest metric values. This suggests that the combinations of height, religious affiliation, age, sex, marital status, creatinine level, family history, temperature, and type of disease, jointly predict the systolic blood pressure more than the other variables. Furthermore, the model on systolic blood pressure was more efficient than that of diastolic BP for the hypertension case in this study while the model on urea was more efficient than that of FBS in the type 2 diabetes case. This result was the same with that of the Ridge regression, as expected. On the other hand, LASSO performed better than Ridge for all the models except diastolic BP. The better performance of the LASSO method can be attributed to the fact that it performs variable selection while the Ridge regression does not. The case of diastolic BP where the Ridge was better than LASSO might be an indication that LASSO has excluded one or more important variable(s) from the final model.

The result in Table 2 suggests that height and each of the four dependent variables are positively related, similarly for age and creatinine. In other words, increase in each of these variables is related to increase in the hypertension and diabetes measures. The negative relationship between marital status and the dependent variables implies that married people are more affected than singles. Recalling the measurement of disease type (X_{13}), the result implies that as we move from the first group (all disease) to other groups, the measures of hypertension and diabetes reduces. More specifically on this, one may infer that those patients with all the diseases stated in group 1 (Renal disease, hypertension and diabetes) have more hypertension and diabetes measures than those in other groups.

4. Conclusion

Hypertension and type 2 diabetes mellitus are two universal health problems. This perhaps explains why researchers throughout the world have continuously attempted to contribute to knowledge on the subject matter for possible solutions to the prevalence of the diseases and their attendant morbidity and mortality. In this research, attempt was made at identifying important variables that could be used in building statistical models to complement the already existing medical structure on the detection of hypertension and type 2 diabetes mellitus. Some statistical and machine learning tools were utilized to estimate the parameters in order to arrive at the models, which were validated by two error performance metrics. Two methods were compared; the one with the capacity for variable selection (LASSO) and the other (Ridge regression) that is characterized with returning all variables into the final model. The result showed that LASSO performed better than Ridge regression in this case. The initial thirteen explanatory variables were reduced to nine (height, religion, age, sex, marital status, creatinine, family history, temperature, and type of disease) by

the process of the LASSO model. The four models, systolic BP, diastolic BP, FBS and Urea, have the same set of explanatory variables for which comparison was made to ascertain the one with the best performance. The model on systolic BP had the best fit based on both MSE and RMSE criteria. This result suggests that, while the set of explanatory variables jointly predict each one of a patient's fasting blood sugar, Urea, systolic and diastolic blood pressure, they are most suitable for predicting systolic blood pressure which measures hypertension.
On the statistical implications of this study, the multicollinearity status of the explanatory variables implied that the analysis would not be possible with OLS while the regularization approaches (LASSO and Ridge) were able to handle it. Identifying and including other explanatory variables such as smoking, excess salt intake, race/ethnicity not considered in this work might improve the prediction performance of these models and provide more information on each of the dependent variables. Future work will consider other estimation methods that are robust to multicollinearity and make comparisons across various scenarios.

Acknowledgment

The authors appreciate the valuable suggestions of the anonymous reviewers in improving the quality and content of the manuscript.

References

Amanda, E. C. (2020). Identifying Influential Variables in the Prediction of Type 2 Diabetes Using Machine Learning Methods. Public Health Thesis, Georgia State University, https://scholarworks.gsu.edu/

Anna, L. and Monika, L. (2005). Marital Difference in Blood Pressure and Risk of Hypertension among Polish Men. European Journal of Epidemiology, 20, 421-427.

Ateeq, M. K., Sharathchandra, R. G., and Meenakshi, R. (2019). Gene Prediction in Heterogeneous Cancer Tissues and Establishment of Least Absolute Shrinking and Selection Operator model of lung squamous cell carcinoma. International Journal of Life Science and Pharma Research, 9(4), 34-8.

Cappuccio, F. P., Miller, M. A. (2016). Cardiovascular Disease and hypertension in Sub-Sahara Africa: Burden, Risk, and Intervention. Internal and Emergency Medicine, 11(3), 299-305. DOI: 10.1007/s11739-016-1423-9

Chai T. and Draxler R.R. (2014). Root Mean Square Error (RMSE) or Absolute Error (MAE)?- Arguments against avoiding RMSE in the Literature. Geoscientific Model Development Discuss,7, 1247 – 1250. https://doi.org/10.5194/gmd-7-1247-2014

Chataut, J., Adhikari, R. K. and Sinha, N. P. (2011). Prevalence and Risk Factors for Hypertension in Adults Living in Central Development Region of Nepal. Kathmandu University Medical Journal, 9(33), 13-18.

Chinghway, L. and Bin, Y. (2015). Estimation Stability with Cross Validation (ESCV). Journal of Computational and graphical Statistics, 25(2), 464-492 https://doi.org 10.1080/10618600.2015.1020159

Davidson, M. B. (2001) "How Do We Diagnose Diabetes and Measure Blood Glucose?" Diabetes Spectrum, 14(2), 67-71. https://doi.org/10.2337/diaspect.14.2.71

Emmert-Streib, F. and Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. Mach. Learn. Knowl. Extr., 1(1), 359-383. https://doi.org/10.3390/make1010021

Hilawe, E. H., Yatsuya, H., Kawaguchi, L., and Aoyama, A. (2013). Differences By Sex in The Prevalence of Diabetes Mellitus, Impaired Fasting Glycaemia and Impaired Glucose Tolerance in Sub-Saharan Africa: A Systematic Review and

Meta-Analysis. Bulletin of the World Health Organization, 91(9), 671-682 DOI: 10.2471/BLT.12.113415

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Non-orthogonal Problems. Technometrics, 12, 55-67.

Hu, F., and Zhang, T. (2020). Study on Risk Factors of Diabetic Nephropathy in Obese Patients with Type 2 Diabetes Mellitus. International Journal of General Medicine, 13, 351-360. DOI: 10.2147/IJGM.S255858

Kearney, P. M., Megan, W., Kristi, R., Paul, M., Paulk, W., and Jiang, H. (2005). Global Burden of Hypertension: Analysis of Worldwide Data. The Lancet Journal, 21, 217-223. DOI: 10,10(6) SO140-b736(05)17741-1

Michel, E. S., Jean-Barthélémy, G., Sola, A. B., Alexandra, Y. and Frédérique, T. (2017). Longitudinal Study of Hypertensive Subjects with Type 2 Diabetes Mellitus Overall and Cardiovascular Risk. Hypertension, 69, 1029-1035. https://doi.org/10.1161/HYPERTENSIONAHA.116.08962

Milani, M. R. M., Andreas, H., Elham, R., and Angelika, P. (2016). Applying Least Absolute Shrinkage Selection Operator and Akaike Information Criterion Analysis to Find the Best Multiple Linear Regression Models between Climate Indices and Components of Cow's Milk. Foods, 5(3). DOI: 10.3390/foods5030052

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramezankhani, A., Azizi, F. and Hadaegh, F. (2019). Association of Marital Status with Diabetes, Hypertension, Cardiovascular Diseases and All-cause Mortality: A Long-term Follow-up Study. PloS One, 14(4), e0215593. DOI: 10.1371/journal.pone.0215593

Saebom, J., Ji-yeon, S., Jaeyong Y., Taesung P., and Mira P. (2019). Structural Equation Modeling For Hypertension and Type 2 Diabetes Based on Multiple SNPs And Multiple Phenotypes. PLoS One, 14(9), e0217189. DOI: 10.1371/journal.pone.0217189

Shen, X., and Huang, H. (2010). Grouping Pursuit through A Regularization Solution Surface. Journal of the American Stat. Assoc., 105 (490), 727-739.

Shizukiyo, I., Kario, K. Kazunori, K., Tadao, G., Naoki, N., Yosikazu, N., Akizumi, T., Eiji, K., and Jichi Medical School Cohort Study Group (2007). Linear Relationship between Blood Pressure and Stroke: The Jichi Medical School Cohort Study. Journal of Clinical Hypertension, 9(9), 677-683. DOI: 10.1111/j.1524-6175.2007.07102.x

Taylor, B. A. (2015). Introduction to LASSO Regression. Yale Statistics.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society Series B (Methodological), 58(1), 267–288.

Tuoyire, D. A. and Ayeteh, H. (2018). Gender Differences in the Association between Marital Status and Hypertension in Ghana. Journal of Biosocial Science, 51(3), 313-334 DOI: 10.1017/s0021932018000147

World Health Organisations (2019). Hypertension. https://www.who.int/news-room/fact-sheets/detail/hypertension. Accessed on 5th July, 2021.

Record Office of the College of Medicine, Ladoke Akintola University of Technology, Osogbo

Appendix

Table 1: Data on Type 2 Diabetes Mellitus and Hypertension

Patient ID	Weight	Height	Religion	Age	Sex	Marital	Creatinine	Dizziness	Headache	Family	Temp	BMI	FBS	UREA	SYS	DIAS	TYPE
1	40	1.55	1	16	1	1	96	2	2	2	36.8	16.65	4.4	2.7	126	65	5
2	41	1.56	2	27	2	2	83	1	1	2	37.1	16.85	3.9	2.2	120	70	1
3	29	1.32	1	21	2	1	96	1	1	2	35.6	16.64	5.1	3.4	110	70	1
4	39	1.66	2	43	2	2	65	1	1	2	38.3	14.15	5.1	4.5	120	74	1
5	24	1.37	1	23	1	1	82	2	1	2	36.3	12.79	9.2	5.8	130	100	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
230	63	1.68	1	79	2	2	110	2	2	2	36.1	22.32	21.9	12.1	150	90	6
231	76	1.59	2	78	1	2	166	1	1	2	39.5	30.06	21.1	4.4	130	70	7
232	61	1.63	1	65	2	2	75	1	1	2	37.3	23	11.2	6.6	150	100	3
233	68	1.61	1	48	2	2	101	2	2	2	35.2	26.23	15.4	6.7	130	70	3
234	51	1.6	2	40	1	1	96	2	2	2	36.8	19.92	21.9	12.1	150	100	5

Table 2: Standardized Type 2 Diabetes and Hypertension Data

Patient ID	FBS	UREA	SYS	DIAS	Weight	Height	Age	Temperature	BMI
1	0.74	0.76	0.79	1.32	1.28	0.51	1.95	0.18	1.16
2	0.79	0.81	0.95	1.08	1.21	0.41	1.39	0.17	1.12
3	0.67	0.69	1.23	1.08	2.01	2.73	-1.7	1.56	1.16
4	0.67	0.59	0.95	0.88	1.34	0.55	0.56	1.55	1.64
5	0.27	0.47	0.68	0.41	2.35	2.25	1.59	0.75	1.89
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
233	0.33	0.39	0.68	1.08	0.61	0.07	-0.3	2.02	0.66
234	0.96	0.12	0.12	0.41	0.54	0.03	0.72	0.18	0.54