Adopting Elastic Net Penalization in Logistic Regression to Achieve Stability: A Case Study of University of UMTHBorno State Nigeria

A. M. Baba[1], S. Maibulangu[2], A. Bishir[3]

[1.2.3] Mathematical Sciences Programme. Abubakar Tafawa Balewa University Bauchi Nigeria

Abstract: ***In this work, study is made on those factors that raise the risk of diabetic patients to be hypertensive and the methods have been determined that best fit the data for modelling. The ordinary logistic regression model fitted turned up to be significant at 1% with a p- value of 0.000. The significant covariates reveal that age increases the risk of being hypertensive for diabetic patient at 16% per unit change. Weight by 60%, marital status decreases it by 74%, exercising decreases it by 95% and family history also by 59%. Various levels of penalization are used for the elastic net logistic regression have shown reasonable improvement over the ordinary logistic regression, with lasso having least deviance ratio (0.52) but also shrinks eight covariates to zero. This is followed by elastic net with centred penalization, which behaves in a similar manner as lasso, but shrinks six of the covariates to zero (with a deviance ratio of 0.58). Ridge logistic regression and the fixed ridge logistic elastic net both retained all the covariates in the model with values of deviance ratio 0.6 and 0.65 respectively.***

Keywords: Logistic regression, Multicollinearity, Ridge, LASSO, Elastic net regression.

1. Introduction

Zero or no dependence among the explanatory variables is one of the assumptions of classical linear regression model. The subject of multicollinearity is widely referred to as the situation where there is either exact or approximately exact linear relationship among the explanatory variables [1]. Multicollinearity can cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests [2] and [3]

Penalized regression methods have received much attention over the past few years, as a proper way to get sparse models in fields with large datasets. The different approaches deal with several issues (e.g. high correlations) in many ways.

ℓ_1 is defined as $\sum_{j=1}^{p}\left|\beta_j\right|$, where β_j is the jth regression parameter, and the summation is taken over p parameters without the intercept.

ℓ_2 is defined as $\sum_{j=1}^{p}\beta_j{}^2$,

with β_j as defined in ℓ_1 above. [3] defined ℓ_2 and named it Ridge. [4] defined ℓ_1 and named it Least Absolute Shrinkage and Selection Operator (LASSO).

A ridge estimator originally developed for linear regression provides a way to deal with the problems caused by multicollinearity. The ridge estimator in general shrinks estimates towards the origin. The amount of shrinkage is controlled by the ridge parameter, whose size depends on the number of covariates and the magnitude of collinearity. The mean squared error (MSE) is guaranteed to be reduced accordingly by the introduction of ridge parameter [3]. [5] applied the ridge regression method to logistic regression to improve parameter estimates and decrease prediction errors. [4] was originally proposed for linear regression models, and subsequently adapted to the logistic case [6] and [7]. Lasso applies an ℓ_1 penalization that, as opposed to ridge regression [3], gives rise to sparse models, ruling out the influence of most of the covariates on the response. Elastic net [8], uses a linear combination of ℓ_1 and ℓ_2 penalties. The elastic net was proposed as a solution to some of the limitations of the lasso, namely the random selection in blocks of highly correlated covariates.

Logistic regression (Logit) is a widely used method for categorical response data. A typical area of application is biomedical studies. A good example is the investigation of the occurrence of a disease (yes/no) as related to different characteristics of the patients.

2. Aim and Objectives

The aim of this dissertation is to adopt the elastic net penalization to achieve stability, by varying and fixing penalty in logistic regression.

Objectives of the study are:

1. To determine the extent of association among the factors that contributes to the development of the disease.
2. Determine the most efficient method of the penalization (ridge, lasso, elastic net) with logistic regression for fitting the data.

2.1 Statement of the Problem

Multicollinearity is a problem in OLS which causes instable estimates and inaccurate variances which affects hypotheses. Ridge regression which adopts ℓ_2 penalization performs better than OLS but yet leaves all the covariates in the model

making it not parsimonious enough. Lasso, which adopts ℓ_1 penalization does well but delete some of the covariate in the model irrespective of its relevance, ones picked one it is correlated with. Elastic net on the other hand combines ℓ_1 and ℓ_2, is said to be more promising than ridge and Lasso by combining good properties of both.

Ridge and Lasso regression methods have been employed in logistic regression and show improvement in performance. However logistic regression models tend to over-fit the learning sample when the number *p* of features, or input variables, largely exceeds the number *n* of samples [9].This work adopts the elastic net, one with varying penalization on both and, with fixed penalization on ℓ_2 norm and varying ℓ_1 norm on logistic regression and assesses the performance with that of logistic ridge and logistic Lasso regressions.

2.2 Significance of the Study

Achieving stability is the motivation behind penalization. The other two frequently used methods (ridge and regression), each has its shortcomings, in which the elastic net performs better by combining the good features of each of the two methods. In logistic regression, various authors adopted the penalization to achieve stability and parsimonious models. This work will serve to compare and make it easy for other adopters of the elastic net penalization in logistic regression to adopt, at a glimpse, the better method of the elastic net in logistic regression (with varied penalization or fixed other). If found to be better, the adoption of the fixed other variable (fixed ridge penalty) gives an easier way of handling the penalized logistic regression.

This study will also go a long way in bringing to light, the risk level at which a patient that is diabetic is likely to be hypertensive; and factors, especially controllable, that can worsen the ailment; and measures to be taken to avoid (if possible) or control the occurrence of the disease based on the results of the study. And also for those that are at risk of the disease to take corrective measures to avoid its menace. This study can also serve as a measure for the general public to read their risk level of the disease. Its can serve as a measure for the government to come up with measures as to how to prevent and control the prevalence of the disease. For those nongovernmental organizations such as World Health Organization (WHO) may use it to come up with enlightenment lectures to reduce it risks of occurrence and, sensitize and educate those diagnosed with the disease on measures so as to reduce the risk of early death.

3. Materials and Methods

3.1 Source of Data

The data for this work is a secondary data, collected from medical records unit of the University of Maiduguri Teaching Hospital, Maiduguri Borno state. The basic information on the individual patients have been collected on the hospital cards by the registry unit of the hospital for easy identification of the patients' files. Variables collected for patients that are diabetic are: hypertension status (which stands as the dependent variable (0/1)), sex, age, marital status, state of origin, current state of living, weight, height, family history of the disease (diabetes), exercising status, smoking status, alcohol intake, salt intake, awareness of diabetes/hypertension relationship and corrective measures. The data collected was for 582 diabetic inpatients for the periods of 2005 to 2009. For each of the patients, the above mentioned variables were collected from their files as captured by the consultant/doctor.

In the data for this work (y) gives the hypertensive status of diabetic patient. The code 1 for hypertensive and 0 for non-hypertensive.(x_1) gives the age of the patient in years.(x_2) gives the weight in kilogramme.(x_3) gives the height. Age, weight and height are already continuous in nature, and so are maintained that way. (x_4) is the body mass index (BMI), which is a measure of overweight, and computed as $BMI = \left(w/h^2\right)$kg/m^2, where w is the weight in kg and h height in meters. (x_5) is thesex of the patient which is coded 1 for male and 0 for female. (x_6) gives the marital status which is also coded 1, 2, and 3, for single, married and divided respectively.(x_7) and (x_8) give the state of origin and state of living respectively, and are coded 1, 2, 3, 4, 5, and 6 are respectively for the six geo-political zones. 1 for north-east, 2 for north-west, 3 for north-central, 4 for south-east, 5 for south-south, 6 for south-west. (x_9) , (x_{10}), (x_{11}), (x_{12}), (x_{13}) and (x_{14}) are respectively for family history, exercising status, alcohol intake, smoking status, salt intake and awareness of diabetes/hypertension relationship and corrective measures, which are coded 1 for affirmative and 0 for non-affirmative.

3.2 Methods

3.2.1 Testing Data for Collinearity

The covariates, $x_i, i = 1, 2, ..., 14,$ for the diabetic/hypertensive patients will all be centred, $x_{ij} - \overline{x}_i$ and for y_i as $y_i - \overline{y}$, such that:

$$\sum_i x_i = 0, \sum_i x_{ij}{}^2 = 1 \text{and} \sum_i y_i = 0.$$

In matrix $X^T X$ all the variables x_{ij} (already centred) are rescaled to have equal lengths by replacing x_{ij} with

$\frac{x_{ij}}{s_i}$, where s_i is the sample standard deviation of the ith variable. The covariates are first subjected for test of multicollinearity using condition number of the of $X^T X$. This is obtained by computing the ratio of the largest to the smallest nonzero singular value, $k = \frac{\sigma_1}{\sigma_r}$, of X. If k is large X is said to be ill-conditioned, where elements of the matrix X are rescaled.

3.2.2 Fitting Logistic Regression

The data collected for the research will be fitted and the various parameters will be obtained using the ordinary logistic equation. The update formula for the likelihood estimator is given by:

$$\beta^{(k+1)} = \beta^k + \left\{ X^T diag\left\{ \pi_i^{(k)}(1-\pi_i^{(k)}) \right\}^{-1} (y - \pi^{(k)}) \right\}$$

When implemented with the iterative reweighted least squares the redefined response at each intermediate step would be

$$z = X\hat{\beta}^k + \left\{ diag\left\{ \pi_i^{(k)}(1-\pi_i^{(k)}) \right\}^{-1} (y - \pi^{(k)}) \right\}$$

With components at

$$z_i^{(k)} = \log \frac{\pi_i^{(k)}}{1-\pi_i^{(k)}} + \frac{y_i - \pi_i^{(k)}}{\pi_i^{(k)}(1-\pi_i^{(k)})}.$$

The penalized logistic regression will be fitted by the formula,

$$\beta^{(k+1)} = \beta^k + \left\{ X^T diag\left\{ \pi_i^{(k)}(1-\pi_i^{(k)}) \right\}^{-1} (y - \pi^{(k)}) \right\} + P_\lambda(\beta^k)$$

using the norm form. Where

$$P_\lambda(\beta^k) = \lambda_1 \left\| \beta^k \right\| + \lambda_2 \left\| \beta^k \right\|^2,$$

for the fixed value of the ridge parameter and

$$P_\lambda(\beta^k) = (1-\alpha) \left\| \beta^k \right\| + \alpha \left\| \beta^k \right\|^2,$$

$$0 \le \alpha \le 1 \quad (3.1)$$

The values of α are varied along the two penalties, i.e. ridge and lasso over the grid of values, 0.00, to 1.00 at a regular interval of 0.05. For each value of α, the data will be fitted and the value of the mean square error will be calculated. Similarly, the other parts of the estimation will be made using the other penalizations. But on the fixed ridge, only the parameter on the lasso part is varied while that of the ridge is fixed with a small value of penalization, say 0.01. This penalization serves to control the issue of high correlation while that of the lasso does the selection. Hence the penalization becomes $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$, with λ_2 fixed and λ_1 varied. In this case too, the $ESS(\beta)$ (the deviance) will be calculated for various values of λ_1. The means squares error of the parameters for the corresponding lasso penalization will be compared for the varied and fixed penalization.

The penalization is done without the slope, β_0, and it's value is obtained thus:

$$\beta_0 = \bar{y}.$$

The deviance ratio is given by:

$$1 - \log(L_k) \Big/ \log(L_0).$$

Where, L_k is the log likelihood of the model with k parameters and L_0 is the log likehood of the empty model.

4. Results and Discussion

4.1 Introduction

In this chapter the results obtained using the methods described in chapter 3, and the statistical packages R2.3.1 and R2.3.2 are presented and discussed.

Table 4.1: Result of the analysis of the hypertensive/ diabetic Patient using Ordinary Logistic Regression

Coefficient	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
x1	0.14646	0.02812	5.209	1.90e-07
x2	0.47227	0.19455	2.427	0.0152
x3	-39.01557	16.08848	-2.425	0.0153
x4	-0.71866	0.45879	-1.566	0.1172
x5	0.30356	0.51651	0.588	0.5567
x6	-1.32621	0.32471	-4.084	4.42e-05
x7	0.54406	0.37727	1.442	0.1493
x8	1.79271	906.88209	0.002	0.9984
x9	-0.89053	0.47768	-1.864	0.0623
x10	-2.94689	1.22327	-2.409	0.0160
x11	0.10382	0.59473	0.175	0.8614
x12	20.81424	1254.06115	0.017	0.9868
x13	0.78045	0.55117	1.416	0.1568
x14	-0.83351	0.64164	-1.299	0.1939

Table 4.1 above shows the result of the analysis of the hypertensive/ diabetic data of the University of Maiduguri Teaching Hospital, Maiduguri. Ordinary Logistic Regression is adopted to fit the model. The model happened to be significant at 1% level with a p-value of 0.0000. The test for significance for the individual parameters reveals that: variables x_1 (age) and x_6 (marital status) are significant at 1% level with p-values of 0.0000 and 0.0004 respectively. x_2 (weight), x_3 (height) and x_{10} (exercising) are all significant at 5% level with p-values of 0.0152, 0.0153and 0.0160 respectively. x_9 (family history) is the only variable that is significant at 10% level, with a p-value of 0.0623. The remaining variables, x_4 (BMI), x_5 (sex), x_7 (state of origin), x_8 (state of living), x_{11} (smoking status), x_{12}

(alcohol intake), x_{13} (salt intake) and awareness of corrective measures are all not significant.

For every diabetic patient, a unit change in age increases the odds of being hypertensive by 16%, unit change in weight increases the odds by 60%, a unit change in height decreases the odds completely (not meaningful), a unit change in marital status, decreases the odds by 74%, exercising decreases the odds by 95% and lastly, family history decreases the odds by 59%.

4.2 Variance Inflation Factor

Table 4.2: Displays the Variance Inflation Factor (VIF)

Coefficient	*VIF*	*Eigen value*
x1	2.003634	3.356817978
x2	52.027540	1.843870802
x3	25.909083	1.794445838
x4	32.249273	1.424268268
x5	1.458843	1.239824050
x6	1.771249	1.058514100
x7	2.442339	0.937964845
x8	1.997712	0.788319525
x9	1.706781	0.656122257
x10	1.455805	0.574080380
x11	1.404446	0.446937281
x12	1.694727	0.349163803
x13	1.354124	0.314998317
x14	1.286545	0.205464405
Y	2.432002	0.009208152
Condition Number	364.5485	>15
Sum(1/eigenvalue)	131.1941	>75

Table 4.2 displays the variance inflation factor (VIF), suggesting presence of collinearity with a condition number of 364.5485, which is by far greater than 15. This indulged the use of the penalized method to the model.

4.3 Logit Models for Ridge and Lasso

Table 4.3: Fitted covariates of logistic ridge regression

Covariates	*Coefficients*
x1	0.05724432
x2	0.03645084
x3	-0.67705060
x4	0.13974018
x5	-0.03074549
x6	-1.26765213
x7	0.15497303
x8	0.02725584
x9	-0.74256350
x10	-0.66950505
x11	0.18019129
x12	1.05984467
x13	0.39090988
x14	-0.18209750

Table 4.3 displays the fitted coefficients of the logistic ridge regression. It can be seen clearly that coefficients are shrunk towards zero, with a deviance ratio value of 0.6, which indicates slight improvement over the ordinary logistic regression.

Table 4.4: Fitted lasso logistic coefficients

Covariate	*Coefficient*
x1	0.04893367
x2	0.01214586
x3	0.00000000
x4	0.16606902
x5	0.00000000
x6	-1.18953713
x7	0.00000000
x8	0.00000000
x9	-0.45041056
x10	0.00000000
x11	0.00000000
x12	0.11210909
x13	0.00000000
x14	0.00000000

Table 4.4 shows the fitted coefficients of the lasso logistic regression selecting only 6 covariates with a deviance ratio of 0.52, smaller than that of ridge logistic regression. The result shows the behaviour of the logistic lasso regression with it style of setting some of the coefficients to zero. The ones picked are 1, 2, 4, 6, 9 and 12.

4.4 Elastic Net with Fixed ridge Penalties

Table 4.5: fitted coefficients of logistic elastic net with fixed ridge penalty

Covariate	*Coefficient*
x1	0.09586184
x2	0.05197878
x3	-2.35519839
x4	0.20377654
x5	0.01411894
x6	-1.39112264
x7	0.28792143
x8	0.02069802
x9	-0.91761587
x10	-1.19785237
x11	0.08718556
x12	2.14066635
x13	0.52292876
x14	-0.30529561

Table 4.5 shows the of fitted coefficients of the logistic elastic net regression with fixed ridge penalty on a value of 0.01.The fitted model gives a deviance ratio of 0.65, slightly greater than that of logistic ridge regression. It can also be observed that all the covariates are retained in the model.

Table 4.6: Fitted coefficients of fitted elastic net with fixed-centred penalty

Covariate	*Coefficient*
x1	0.051252810
x2	0.028787861
x3	0.000000000
x4	0.146594372
x5	0.000000000
x6	-1.201903399
x7	0.000000000
x8	0.000000000
x9	-0.669956034
x10	-0.138923006
x11	0.000000000
x12	0.616883091
x13	0.003480438
x14	0.000000000

Table 4.6 above shows the number of fitted coefficients of the fixed (centred at 0.5) elastic net within a grid of 0.0 to 1.0 at a regular interval of 0.05. Only eight coefficients are fitted. The value of the deviance ratio is 0.58, showing improvement over ridge and varied elastic net. The ones fitted are 1, 2, 4, 6, 9, 10, 12 and 13. The remaining are all set to zero. Table 4.7 gives the values of the respective deviates, indicating decline in the deviance.

4.5 Summary of Fitted coefficients

Table 4.7: Displays the summary of the fitted value as explained above.

Summary of the result of fitted penalized logistic regression

Penalty	*No. fitted*	*Fitted covariates*	*% Deviance*
Logit ridge	all	All	0.60
Logit lasso	6	1, 2, 4, 6, 9, 12	0.52
Logitenet fixed	all	All	0.65
Logitenet centered	8	1,2,4,6,9,10,12,13	0.58

5. Summary and Conclusion

5.1 Summary

Achieving stability in logistic regression using elastic net is the main aim of this thesis. Data on diabetic patients from University of Maiduguri Teaching Hospital (UMTH) was obtained. The variables collected are age, weight, height, sex, marital status, family history of diabetes, exercising status, smoking status, alcohol intake, awareness of corrective measures, and body mass index (BMI) was calculated.

The fitted ordinary logistic model was found to be significant ($p<0.001$), and has significant covariates: age, weight, height, marital status, family history, and exercising status. Variance inflation factor calculated showed that there is multicollinearity. The data set was subjected to the various methods of the penalization with the logistic regression. Ridge method retained all the covariates in the model with shrunk coefficients towards the origin. Lasso on the other hand deleted 8 of the covariates leaving only 6: age, weight, height, marital status, family history and exercising, exactly corresponding to the ordinary logistic regression variables. The ridge has an improved deviance ratio test as compared to the lasso. The elastic net with centred value of penalization performed well, deleting 6 of the covariate and adding alcohol intake and salt intake to the list of lasso, with a better deviance ratio than that of lasso. Fixed value of the penalization behaves exactly like the ridge, leaving all the variables in the model with a deviance ratio value of 0.65.

5.2 Conclusion

From the result obtained in the analysis of the fitted model using the ordinary logistic regression, one can draw the conclusion that an individual that is diabetic is so likely to be hypertensive. This, with the covariates that significantly contribute to the model: age, height, weight, marital status, exercising status and family history of the disease. It could be observed that the lasso logit agrees with the ordinary logistic regression in picking the same variables. Lasso with small fixed ridge, here in this analysis, performs best and only leaving all the variables in the model. It can also be concluded, based on this result, that the centred elastic net logistic regression is better than logistic lasso. However, the ridge on its part makes the realization of the elastic net logistic worth considering. This reveals, in the case where there is large number of covariates (just as in the case of biomedical data or gene study where collinearity is common) logistic elastic net regression will do well in stabilizing the issues of multicollinearity.

References

[1] Gujarati, N. D., Porter, C. D. (2009).*Basic Econometrics*.McGraw Hill International Edition (Fifth Edition).
[2] R. L. Schaefer (1986). *Journal of Statistical Computation and Simulation***25** 75-91
[3] Hoerl, A. and Kennard, R. (1988), Ridge regression, *in 'Encyclopedia of Statistical Sciences*, Wiley, New York, **8**, 129–136.
[4] Tibshirani R. (1996).Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**(1), 267-288.
[5] Le Cessie, S. and Van Houwelingen, J. C. (1992).Ridge estimators in logistic regression, Applied Statistics. *Journal of the Royal Statistical Society, Series C* **41**, 191-201
[6] Roth, V. (2004).The generalized LASSO.*IEEE Transactions on Neural Net-works,* **15**, 16 28.
[7] Shevade S. and Keerthi S (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics,* **19**(17), 2246-2253.
[8] Zou, H. and Hastie T. (2005).Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B,* **67**(2), 301-320.
[9] Zakharov R. and Dupont P., (2011), *machine learning group*, **6,** 133-144.