A Descriptive Model for Risk Factors of Stroke Using LASSO Logistic Regression (Evidence from Ghana)

Adusei Bofa¹, R. K Avuglah² & E. Harris³

¹Adusei Bofa, MPhil., ²R.K Avuglah, PhD. & ³E Harris, MSc.
¹Department of Economics and Statistics, Lecturer, Garden City University College
²,³Department of Mathematics, Senior Lecturer, KNUST
Kumasi, Ghana

Abstract: Using statistically conventional adjustments of existing methods we derived a risk score of stroke. With the least absolute shrinkage and selection operator (LASSO) Logistic Regression we selected the most informative risk factors of stroke as well as established the association with stroke. A sample size of 5119 on the aging population of Ghana as secondary data was used. R statistical software for windows version 3.1.1 with the penalised package was used in implementing the LASSO-based Logistic Regression. Hypertension is a crucial risk factor for stroke. Concerning developing stroke an individual who exercises repeatedly had a little risk of getting stroke compared to a person with no exercise. Based on the data about stroke we discovered angina and arthritis as emerging risk factors of stroke in addition to hypertension, physical activities, age, and diabetes. Regarding stroke cases in Ghana, We hypothesize that non-existence of physical activities, hypertension, and diabetes reflect the trend of stroke cases. There is a need for local awareness about the importance of hypertension management in Ghana.

Keywords: Statistically conventional, Risk score of stroke, LASSO, Logistic Regression, Ghana.

1. INTRODUCTION

Concerning cardiovascular illnesses at hand is a group of related risk factors that start to exist in individuals (Anderson et al., 1991). To assess the natural settings of cardiovascular infection likelihood with reference to the argument of bringing together the weight of several risk factors is extra exact in all
facts than to set out single risk factors, for the motive that the collective special effects of many factors may be synergistic as the average reduction in risk factors of a person is significant and possibly will yield an anticipated result than a single risk factor reduction (Jackson et al., 2005).

In Ghana, the effect of persistent infection in the midst of the aged masses as well as a probable result on the living conditions of advanced years adults populace set up a significant challenge. These Non-Communicable Diseases (NCD) threats have a consequence on the varying behavioural forms at an astonishing price to health and over the economy (Bloom et al., 2011). Regarding stroke patient's prediction statistical models with simple risk elements or factors may accomplish similar to expert clinician outcome.

Scholars latch onto their doings to find risk factors or covariates for the dissimilar types of cardiovascular diseases. To put forward an insight on which elements or factors among the cluster of potential risk factors (predictors) that are strongly related or most informative about stroke is a vital demand that still necessitates additional solution. Mathers and Loncar (2005) forecast that in the successive two decades stroke will keep on among the main four ordered roots of mortality with reference to developing countries. In modern times it has turned out to be very challenging by means of low bias and low variance model to determine the most informative risk factors of stroke from the potential risk factors. Hence the use of LASSO regression.

The aims of this work were to use LASSO Logistic Regression (LLR) to select the most informative risk factors of stroke and to come up with a risk score of stroke with statistically acceptable adjustments of current methods. To lessen the load of untimely demise and ill health, augmented research on stroke is therefore necessary to rally more effective recommendations for avoiding stroke.

As to the analysis of data on stroke, it was done in two folds: The inferential statistics focused on LLR, predicted probabilities, odds ratios derivation, cross validation of model and parameter estimation using R statistical software for windows version 3.1.1. Secondly, we used the Statistical Product and Service Solution (SPSS: IBM version 20) for the descriptive analyses. The agreement of suggested techniques of weight application was completed by way of Bland and Altman test with R software and internal consistency was checked using Cronbach's alpha.

2. MATERIALS AND METHODS

Data: Concerning data for this work a secondary data known as Sage 1 form World Health Organization's (WHO) was used which is detailed formerly in Kowal et al. (2012). Using a sample of size 5119, the variables comprised Lifestyle: smoking alcohol intake, diets record, and physical activities. Chronic illness collected were hypertension, asthma, arthritis, diabetes, and angina. Classification variables: education level, marital status, location (urban or rural), ethnicity, gender, and age. The binary dependent variable used was the absence or presence of a stroke.

The model description LASSO Logistic Regression (LLR): The LASSO is a selection and shrinkage technique designed for regression model (Tibshirani, 1996) initially applied to Ordinary Logistic Regression. For a better explanation of model and identification of those factors (risk factors) which remain utmost strongly linked to the outcome, the LASSO put some constraints on the model parameters with a specified constant as an upper bound to the sum shrinking some parameters towards zero (0). Equation 1 shows the ordinary Logistic and the join Log likelihood for \( \beta \) is represented labelled equation 2.

\[
\log \left[ \frac{\pi_i}{1 - \pi_i} \right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad (1)
\]

\[
l(\beta) = \ln(\beta / y_{i_1} \ldots y_n) = -\sum_{i=1}^{n} \left[ (1 - y_i)\beta^TX_i + \ln(1 + \exp(-\beta^TX_i)) \right] \quad (2)
\]

With reference to equation 1 and 2, \( X_{ij} \) is the complete collection of explanatory variables. By adding the LASSO constraint to the joint-likelihood in the 2 it yielded the constraint equation below.
\[ l(\beta^{\text{lasso}}) = -\sum_{i=1}^{n} [(1 - y_i)\beta^T X_i + \ln(1 + \exp(-\beta^T X_i))] + \lambda \sum_{k=1}^{p} |\beta_k| \]  

The \( \lambda \) which is a tuning parameter can be selected by the user calculated using the methods: a variant of steins, generalized cross validation and cross validation (Tibshirani, 1996). Applying this \( l_1 \) norm on the ordinary Logistic Regression (OLR) yields the LASSO Logistic Regression (LLR). Some coefficient in the vector \( \beta \) become zero depending on how sufficiently \( \lambda \) is specified.

**Methods of Applying Weight \( (W_i) \):** In risk estimation there exist a wide application of methods; as indicated by the table below

Table 1: Depiction of prevailing techniques that are commonly useful now in medical risk estimate (Vaitheeswaran et al., 2014)

<table>
<thead>
<tr>
<th>Codes for method</th>
<th>Weight</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>([10\beta_i])</td>
<td>((-\infty, \infty))</td>
</tr>
<tr>
<td>M2</td>
<td>((2\left[\frac{\beta_i}{\min({</td>
<td>\beta_i}</td>
</tr>
<tr>
<td>M3</td>
<td>((2\left[\frac{\beta_i}{\text{mean of smallest two } \beta's}\right]))</td>
<td>((-\infty, \infty) - {0})</td>
</tr>
</tbody>
</table>

The standard method (LASSO estimate) determines the risk scores not having any transitional round off for decimal places. The other three methods M1, M2 and M3 take into account the non-zero \( \beta \) coefficient in the LASSO Logistic Regression model. Based on LLR model, this study found the mean of the two absolute largest value of the \( \beta \) to be the denominator as compared to M3. Hence for method M5 100 \( \frac{\beta_i}{\text{mean of largest two } \beta's} \) Bland and Altman test was used to check the level of agreement plus Cronbach’s alpha to determine the readability of methods.

In exploring the data, of all 5119 individual contained in the data used for this study, 2714(53\%) existed for male individuals and 2405(47\%). The distribution for age (years) of the subjects is virtually normal (mean = 60 years; median = 60.09 years; skewness = -.172), with a range of 19 years to 120 years. Also, 2107(41\%) of individuals were located in the urban area whereas 3012(58.8\%) were in the rural.

### 3.1 Optimal Shrinkage \( \lambda \) Parameter

K-fold method for cross-validation size was used in obtaining the lambda ($ \lambda $) value by applying the “profL1” function in order to profile the cross-validated log-likelihood for the covariates and factors in the LLR model in which all covariates and factors were penalized. From “profL1” using 10-fold cross validation was found at $ \lambda = 3.410935 $ at 10-folds giving as the global Optimum and that of “optL1” using 10-fold cross-validation was found at $ \lambda = 2.988797 $.
Figure 1 give details about profiling the choice of lambda for LLR model with all risk factors penalized.

The optimum value for $\lambda$ from “optL1” function within the “penalised” package ($\lambda = 2.988797$) depicted with the green short dashed line. The $\lambda$ value obtained to maximize the cross-validation log likelihood using the “profL1” function of the “penalised” package ($\lambda = 3.410935$) is the red short dashed lines.

3.2 The LLR Model for Stroke (All Risk factors Penalised)

In the figure 2, summary of how the estimates of the coefficient for the LLR model of stroke in which all independent variables were penalized change with the shrinkage parameter $\lambda$ was then schemed with the “plotpath” command of the “penalised” package. Beginning with the initial value of $\lambda$ that marks all the parameters shrunk to zero(0), a pre-defined number of models are created with values of $\lambda$ not matching another until the last $\lambda$ is one and the same to that indicated in the set-up of the LLR model.
Figure 2: Coefficient estimates set against $\lambda$ values for LASSO Logistic Regression model of Stoke ($\lambda = 3.410935$)

A number of the coefficients estimates for these risk factors of stroke were fixed to precisely 0 using the LLR as described above by the penalized package with the R output. Such examples comprises chronic lung disease, asthma, rapid walk and alcohol. There were 13 non-zero coefficients with $\lambda$ set to be equal to (3.410935) with cross-validated estimate log likelihood of -493.3347 at a LASSO penalty of 16.73709 and an intercept of 2.02347071. Example on the none-zero coefficients include Hypertension, physical activity (Exercises), diabetes and gender.

Table 2 displays the coefficients labelled LLR point estimate, odds ratio and p-value. The LLR coefficients portrayed by the table gives the variation in the log odds outcome for a one unit increase in the predictor variable (risk factors):

- From the results a person who does extreme exercise has a less log odds of $-0.41595175$ compared with a person who do not exercise. This clarifies why the coefficient is negative from Table 4.1 above. Inferring a person with extreme exercise has a less chance of developing stroke.
- An individual with Angina has a high possibility thus large log odds of $0.03743824$ developing stroke this evident with its positive coefficient.
Table 2: Distribution of risk factors (predictors) and LLR none zero coefficients

<table>
<thead>
<tr>
<th>Risk Factors (Predictors)</th>
<th>LLR Point Estimate</th>
<th>Odds Ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender(0=female;1=male)</td>
<td>0.04413259</td>
<td>1.0451209</td>
<td>0.933</td>
</tr>
<tr>
<td>Location(0=rural;1=urban)</td>
<td>0.27462354</td>
<td>1.3160351</td>
<td>0.000</td>
</tr>
<tr>
<td>Education(0=no formal;1=formal)</td>
<td>-0.05383929</td>
<td>0.9475844</td>
<td>0.130</td>
</tr>
<tr>
<td>Physical Activities(0=no ;1=yes)</td>
<td>-0.41595175</td>
<td>0.6597121</td>
<td>0.000</td>
</tr>
<tr>
<td>Ethnic(0=Akan;1=others)</td>
<td>-0.05333828</td>
<td>0.9480593</td>
<td>0.085</td>
</tr>
<tr>
<td>Fruit intake(0=not regular;1=regular)</td>
<td>-0.11438117</td>
<td>0.8919179</td>
<td>0.102</td>
</tr>
<tr>
<td>Angina(0=no; 1=yes)</td>
<td>0.03743824</td>
<td>1.0381479</td>
<td>0.000</td>
</tr>
<tr>
<td>Arthritis(0=no; 1=yes)</td>
<td>0.06544823</td>
<td>1.0676375</td>
<td>0.004</td>
</tr>
<tr>
<td>Tobacco(0=used; 1=yes)</td>
<td>-0.01857579</td>
<td>0.9815957</td>
<td>0.285</td>
</tr>
<tr>
<td>Diabetes(0=no; 1=yes)</td>
<td>0.21326979</td>
<td>1.2377185</td>
<td>0.000</td>
</tr>
<tr>
<td>Hypertension(0=no; 1=yes)</td>
<td>0.22971266</td>
<td>1.2582384</td>
<td>0.000</td>
</tr>
<tr>
<td>Age(0=18-55; 1=56+)</td>
<td>0.17412797</td>
<td>1.1902079</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The column in table 3 label odd ratio gives the coefficients as odds ratios. Odds ratio which is obtained by taking the exponent of the coefficient, can be inferred as the multiplicative transformation in the odds for a one unit variation in the predictor variable (risk factors).

The risk factors that the LLR selected are; gender, location, education level, physical activities(exercise), ethnic, fruit intake, angina, arthritis, tobacco usage, diabetes, hypertension and age. The most informative or causative risk factors of stroke thus significant at alpha level of 0.05 are; diabetes, hypertension, physical activities (exercise), angina, arthritis and location of an individual. Whereas gender, ethnic, fruit intake and tobacco were not significant causative of stroke.

**Figure 3** shows the predicted probability of individual’s base on the LLR model as against the age of individuals.
The LLR model predicted that if a person is in the range of 18-53 the chances of developing stroke is 0.9 (90%) this may as a result of some lifestyle concerning that age group, depicted by the area before the vertical red line. The whole population is at risk with at least 70% likelihood.

4. SCORE ALLOCATION TO RISK FACTORS

Below gives the scores assigned by the methods of weight application to risk factors whose estimate are non-zero and significant. The weight selection for second method (M2) selected the regression coefficient (0.01857579) as the absolute minimum, while third method (M3) computed the score by the mean (0.02800701) of the absolute value of the two smallest regression coefficients. Again fourth method (M4) selected the absolute maximum regression coefficient of (0.41595175) and fifth method (M5) selected the mean (0.3452876) of the absolute value of the two largest regression coefficient.

The table 3 below depict the score given by the three existing methods and the extra two proposed by this study detailed in method and material section. Concerning the risk score of an individual: if any of the risk factor indicated in the table is exiting, then the score allocated to that risk factor is recorded. The process will continue till all scores are recorded. The total recorded score specifies how much that individual is at risk.

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Weighting Methods and Respective Scores Allocated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SM</td>
</tr>
<tr>
<td>Location (Urban)</td>
<td>0.27462354</td>
</tr>
<tr>
<td>Exercise (Yes)</td>
<td>-0.41595175</td>
</tr>
<tr>
<td>Angina (Yes)</td>
<td>0.03743824</td>
</tr>
<tr>
<td>Arthritis (Yes)</td>
<td>0.06544823</td>
</tr>
<tr>
<td>Diabetes (Yes)</td>
<td>0.21326979</td>
</tr>
<tr>
<td>Hypertension (Yes)</td>
<td>0.22971266</td>
</tr>
<tr>
<td>Age (55+)</td>
<td>0.17412797</td>
</tr>
</tbody>
</table>

Figure 4 depict the Bland and Altman test. The space between the mean line (-11.49359) and the parallel green line corresponding to zero on the x-axis indicates the bias between the two methods (M4 and Sm). The upper limit's (114.90989) confidence interval was (62.63143, 167.18835) and that of the lower (-137.89707) was (-190.17553, -85.61861).

Since the line of equality (0) on the x-axis is within the error margin (-41.67657, 18.68940) of the mean difference (-11.49359) line implies that the bias is not significant and hence there is a level of agreement between the two methods. Again the bias was not significant implying a level of agreement for the two methods.
5. RELIABILITY MEASURE FOR SCORING METHODS

The table 5 below specifies the level of reliability of methods with the standard method. The method label M3 has the lowest Cronbach's alpha value and Intra class correlation of 0.793 and 0.004 respectively. Whereas M1 had the highest value of Cronbach's alpha of 0.812 and an intra class correlation of 0.198 at a significant p-value (0.00) indicating a good internal consistency between M4 and standard method thus the weight given by the LLR estimate.

Table 4: Measure of internal consistency among weight application methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Cronbach's alpha</th>
<th>Intra class correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point Estimate</td>
<td>95% of confidence interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lower Bound</td>
</tr>
<tr>
<td>M1</td>
<td>.912</td>
<td>0.198</td>
</tr>
<tr>
<td>M2</td>
<td>.818</td>
<td>0.003</td>
</tr>
<tr>
<td>M3</td>
<td>.793</td>
<td>0.004</td>
</tr>
<tr>
<td>M4</td>
<td>.826</td>
<td>0.007</td>
</tr>
<tr>
<td>M5</td>
<td>.798</td>
<td>0.005</td>
</tr>
</tbody>
</table>
6. CONCLUSION
Data about stroke discovered angina and arthritis as an emerging risk factor of stroke in addition to hypertension, lack physical activities, aging and diabetes. We hypothesise that hypertension, diabetes and lack of physical activities reflect the trend of stroke cases in Ghana.

7. RECOMMENDATIONS
It is a necessity to embark on public consciousness about the emerging risk factor like angina, arthritis and the significance of handling hypertension although stroke is observed as a severe and not inevitable condition in Ghana. Upgraded research, nation-wide and world-wide commitment for the avoidance and control of stroke in Ghana and Africa as a whole is emphasized. There should be a research on why the younger adults are associated with high possibility of developing stroke.

8. REFERENCES
Caster, O., G.N., N., Madigan, D., and Abate (2008,). Large-scale regression based pattern discovery in international adverse drug reaction surveillance., Proceedings of the KDD Workshop on Mining Medical Data and KDD Cup.


Zambrut Journal, Link Access;
https://zambrut.com
https://zambrut.com/risk-stroke/

© Copyright International Journal of Zambrut | Zambrut, Inc.