KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

A DESCRIPTIVE RISK MODEL FOR RISK FACTORS OF STROKE USING LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR

By

Adusei Bofa (Bsc Statistics)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MPHIL MATHEMATICAL STATISTICS

October 9, 2016

Declaration

I hereby declare that this submission is my own work towards the award of the MPhil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for the award of any other degree of the University, except where due acknowledgement has been made in the text.

Adusei Bofa ....

Student(PG2520514) Signature Date

Certified by:

Dr. R.K Avuglah ....

Supervisor Signature Date

Certified by:

Dr. R.K Avuglah ....

Head of Department Signature Date

Dedication

This work is dedicated to my mother Akua Brago and grandfather, Nana Senyah.

Abstract

We selected most informative risk factors of stroke as well as established the association with stroke and derived a risk score of stroke using statistically acceptable modifications. Secondary data on the aging population of Ghana was used with a sample size of 5119. The LASSO Logistic Regression was implemented through "R"statistical software for windows version.3.1.1 using the penalised package. A key risk factor identified was hypertension. A person who exercises regularly had a low risk of developing stroke as compared to one who did not exercise. Data about stroke discovered angina and arthritis as emerging risk factors of stroke in addition to hypertension, physical activities, age and diabetes. We hypothesise that hypertension, diabetes and lack of physical activities reflect the trend of stroke cases in Ghana. There must be public consciousness about the significance of managing hypertension in Ghana.

Acknowledgements

My sincere and heartfelt appreciation goes to the Almighty God, who has brought me this far.

I also wish to express my profound gratitude to Dr.R.K.Avuglah, my supervisor of the Department of Mathematics KNUST for the tremendous help, advice and direction throughout the writing of this thesis.

I wish to acknowledge all lecturers and staff of the Department of Mathematics for the various roles each one of them played towards the successful execution of this study. Am also thankful to Dr.Akwasi Preko Department of Physics KNUST,Prof.Bashiru Saeed of Kumasi Polytechnic Faculty of Applied Science and the World Health Organisation for making it possible to access Sage Wave 1 .

I extend my appreciations to Kofi Nyampong and Kwaku Nimoh -two individuals who have helped me in means that one cannot begin to mention.

Lastly to my love ones,Stephen Obour, Afia Nyarko,Richcane Amankwa and Saforo Robert Mensah.Thank you for your effort you played.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There is a group of similar risk factors that start to exist in individuals concerning cardiovascular diseases (Anderson et al., 1991).To estimate the nature of cardiovascular disease possibility concerning the argument of bringing together the weight of numerous risk factors is extra correct in all details than to deploy single risk factors, for the reason that the increasing effects of multiple factors may be synergistic since the average reduction in risk factors of an individual is significant and may produce a desirable result than a sole risk factor reduction (Jackson et al., 2005).

The subdivision of medicine anxious by means of the occurrence and spreading of infection and additional issues concerning health condition has sensitized principal care clinicians to the abstract idea of risk factors for disease. Most researchers and clinicians define risk factors as features of individuals used to point out increased likelihood of having or developing a disease . Regretfully many circumstances affecting the existence of cardiovascular disease (Stroke) have more than one risk factor that need to be assessed, there by adding a level of difficulty to the assessment of risk (Jackson et al., 2005). Hence this study employed LASSO logistic to the risk factors of stroke to assess the risk level for the ageing population in Ghana.

1.1 Background

Ghana remains one of Sub- Sahara African republics undergoing epidemiological shift in which prolonged slow to heal diseases are going up especially in the midst of the elderly populace . This concurs with populace ageing experience entirely towards provinces of the domain(DeSA, 2013).

With current general direction in which demographic characteristics are changing generally, prolonged illnesses situated more or less centrally in relation to the elderly is increasing by way of intense implication on behalf of life standards in relation to mature grownups. 2010 Global Burden of Disease Study makes available indications on the shifting developments in disease outlines within sub-Sahara Africa by means of notable large decrease among Disability Adjusted Life years(DALYs)/100 when taken as a whole, on a contrary an excessive upturn among proportion of DALYs being caused by prolonged disorders(stroke) particular associated with to mature grownups. Stroke is the second leading condition that gives rise to death worldwide and to acquired disability in elderly in most regions (Feigin, 2007).

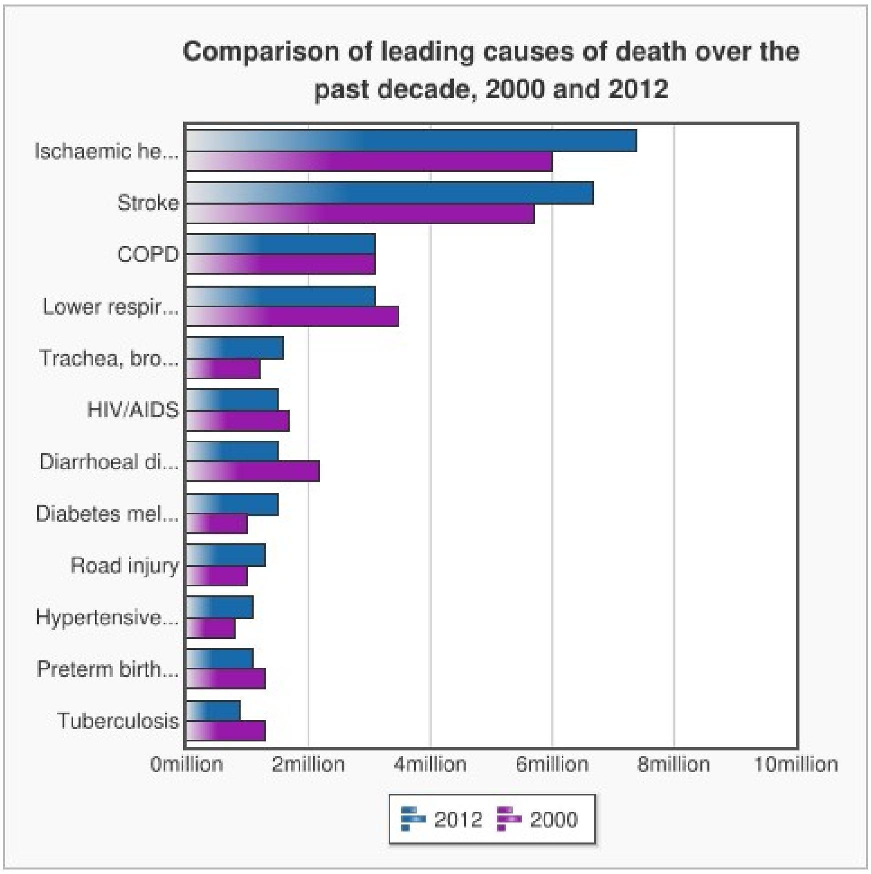


Figure 1.1: The 10 leading causes of death in the world, 2000 and 2012.(WHO,2014.)

In Ghana, the impact of prolonged disease in the midst of the aged populace as well as a likely consequence on the living standard of aged adults populace constitutes a significant challenge. These Non-Communicable Diseases (NCD) risks have effect on the altering behavioural forms at an extraordinary price to health plus economy (Bloom et al., 2011). Statistical models with simple risk factors or variables may accomplish similar to expert clinician outcome in stroke patient's prediction. This study sought to explore the use of Least Absolute shrinkage and Selection Logistic regression to detect stroke risk factors in Ghana.

1.2 Problem Statement

Stroke is a medical condition where blood flow to the brain is cut off.(Schulz et al., 2004). Despite the fact that some clinical pointers of a kind; like Blood Glucose Level (BGL), Blood Pressure (BP) and Haemoglobin of a healthy individual may perhaps not remain alike once the individual is not fit and possibly will form the core control reason for existence or absence of a particular illness in the individual, one cannot overlook the effect of some risk factors (e.g. weight, age, gender). Persons sometimes by their lifestyles formulate grounds for disease to develop. Some of these are alcoholism, smoking, lack of regular exercise, poor eating habit etc.

Researchers persist in their activities to find risk factors for the different kinds of cardiovascular diseases. Nevertheless a vital demand that still necessitates additional solution remains to put forward an insight on which covariates or factors in the cluster of potential risk factors (predictors) that are strongly related or most informative about stroke . Mathers and Loncar (2005) detailed that the comprehensive assessments of disease problem forecast that in the subsequent two decades stroke will remain among the highest four ordered roots of mortality concerning developing countries.

In recent times using low bias and low variance model to determine the most informative risk factors of stroke from the potential risk factors has become very challenging. Hence the application of LASSO regression.

1.3 Purpose of the study

Stroke remains a developing community health problem with a demanding task within Ghana calling for earnest and insistent consideration for the aforementioned control. For the reason that some of the risk factors are modifiable concerning stroke, characterization of these risk covariates or factors among Ghanaian ageing population in relation to effects of stroke are earnestly necessary to direct policies on non-infectious diseases Minicuci et al. (2014).

The Ordinary Least Square estimations are computed by way of minimizing the sum of square error. According to data analyst at hand remain two aims for which they are frequently not content with the OLS estimations. First stays prediction precision: the OLS estimations are more repeatedly associated with large variance but no bias. Sacrificing a little bias the large variance of the predicted values from the OLS method is reduced by penalising or adjusting irrelevant coefficients to zero (0) therefore improving overall prediction accuracy. Secondly, interpretation; we are much concern with a smaller that shows the strongest effects (Tibshirani, 1996).

Preferably, the desire of statisticians is to identify a manageable subsets of risk factors that are found to be associated with stroke. In addition, to improve the prediction accuracy, the numerical instability needs to be minimized. In order to overcome this new found defect of OLR, a constraint form of Logistic regression has been put forward for consideration (Caster et al., 2008) that makes practical and effective use of LASSO technique (Tibshirani, 1996) method. This alternative shrinkage and selection method known as LASSO Logistic Regression (LLR) takes action as an ideal means of solving the aforementioned problem. It is able to shrink parameter estimate getting to 0- in some cases, equate parameter to 0(thus excluding them from the model).

This study is meant to evaluate the association concerning risk covariates of stroke that is most informative and identify emerging risk factor. Model and predict the incidence of stroke using LASSO Logistic Regression centred on World Health Organization's (WHO) multi country survey on world-wide Ageing and Adult Health (SAGE) Wave 1 Ghana data. Details of discoveries on domestic features of the people are found inKowal et al. (2012).

1.4 Research Objective

With collective observations from the problems associated with OLS estimation and available methods of applying weight to risk factors this study aims:

- To explore the application of LASSO Logistic Regression to select the most informative risk factors of stroke to establish association with stroke.
- To derive a risk scoring method for stroke using statistically acceptable modifications of existing methods.

1.5 Structure of Methodology

This work used secondary data centred on World Health Organization's (WHO)2008 multi country survey on world-wide Ageing and Adult Health (SAGE) Wave 1 Ghana with a sample size 5119. Since the dependent variable stroke has a binary outcome and a vector of risk factors, we constructed LASSO logistic.

LASSO is a shrinkage and selection technique aimed at better interpretation of model and identification of those covariates (risk factors) which are intensely related with dependent stroke variable , the LASSO will shrink some parameters towards zero (0) by putting some constraints on the model parameters with a specified constant as an upper bound to the sum (Tibshirani, 1996). The LASSO Logistic regression was implemented by R statistical software for windows version.3.1.1 using the Penalized package and focused on ; parameter estimation, odds ratios derivation, predicted probabilities and cross validation of model.

Before fitting the penalized regression model, we test the global null hypothesis of no association between any of the explanatory variables and the response through globaltest package which ties very closely to the penalized package.

1.6 Significance of the Study

Ample effort remains as the risks and penalties of stroke remain disturbingly high. Stroke must thus principally, be a main concern on health plan in the developing world. Augmented research on stroke is therefore necessary to improve more effective guidelines for avoiding stroke to lessen the load of untimely death and ill health. A study of this form is thus not merely timely but also essential to make available more awareness into some facets of the problem and to make available the base for extra studies and suitable intervention through LASSO logistic regression application.

1.7 Limitation of the Study

One limitation of the study is that the model might lose approximately weighty explanatory or independent variables along the way subject to the degree of the shrinkage parameter . The LASSO penalization have the tendency to select one in a cluster of extremely correlated variables. We will employ group LASSO to mitigate this problem in our next work.

1.8 Organisation of the Study

The 1st chapter describes how the LLR can improve the OLR modelling of stroke and risk factors, the advantages associated with LLR and the implications of LLR with respect to estimating coefficients. Chapter 2 gives information about literature review. Chapter 3 details the methodology of this work. The results of the application of LLR model to the presence or absent of stroke is evaluated, and the optimal shrinkage parameter for the model are all discussed in chapter 4. Again the measure of internal consistency and agreement measure are also found in Chapter 4 . Chapter 5 provide information on summary of findings. References and appendices comes next.

Chapter 2

Literature Review

2.1 Introduction

Traditionally stroke is well-explained as fast emerging medical indicators of central disorder of cerebral function, enduring at least a day or causing demise by way of no readily perceived or understood reason other than that of vascular source Aho et al. (1980). Stroke stands a public neurological problem accounting for a third of all deaths in Western countries Sudlow and CP. (1996). Stated rates of stroke reappearance have change extensively as of 3% to 22% within a year, 10% to 53% by five years in dissimilar studies. (Zumo, 2006).

Capuccio (2004) discussed that stroke and ischaemic heart condition are now the highest found causes of mortality in the world and that 70% of these deaths happen in developing countries, above and beyond the impact of the HIV/AIDS epidemic, and that stroke and ischaemic heart condition will keep on the most found causes of morbidity, disability, and death in developing countries in 2020.

2.2 Diet as a risk factor

Ecological influences, comprising diet, indicate a crucial part in stroke growth. Nutrition of a person is connected to extra lifestyle issues (alcoholism, physical activities, etc.)Rose (1981).There is further indication on the part of overweight in risk adjustment than other nutritional aspects Stamler et al. (1993).

2.2.1 Stroke and Saturated Fat

Cochrane examination of twenty seven test of suitability (18,196 partakers) studied outcome on lessening or change in food fats intended for more than or equal 6 months on lessening serum cholesterol statuses and going on overall cardiovascular death as well as disease. The assessment comprised (seven) 7 test of high suitability, six (6) moderate test of suitability (six) and fourteen (14) little risk partakers. Men comprised extraordinary risk partakers exclusively. Regarding entire death (frequency ratio 0.98, 95% CI 0.86 towards 1.12) at hand were no significant effect, a trend concerning safety guard against cardiovascular death (frequency ratio 0.91, 95% CI 0.77 towards 1.07), essential guard against cardiovascular happenings (frequency ratio 0.84, 95% CI 0.72 towards 0.99). This result was trivial not considering high risk of bias. Trials by means of not less than twofold years of monitoring gave firmly signal on guarding against cardiovascular happenings (frequency ratio 0.76, 95% CI 0.65 towards 0.90). The assessors established a minor but possible significant lessening in relation to cardiovascular hazard with a lessening otherwise change of food fat consumption, comprehended mainly in longer period studies.(Hooper et al., 2004).

2.2.2 Intake of Fruit and Vegetables

Twofold organized reviews of cohort readings observed the advantage of fruit and vegetable consumption for the decline of cardiovascular risk. It was confirmed after numerous cohort readings that relative risk of stroke is reduced as a result of improved vegetable intake (relative risk 0.77) and fruit (relative risk 0.86) consumption in a single assessment also 15% dropped in relative risk concerning cardiovascular disease among persons taken in lots of fruit and vegetables matched to those taking in small (the same as four-fold upturn in fruit and multiplying up vegetables) in one more review Vestfold (2003).

2.3 Stroke and Weight

A methodical appraisal of randomized controlled trial on dietary to decrease weight assessed the consequence on B.P stood known. Insignificant number of patients remained involved (6 trials involving 361 members) (Mulrow et al., 2004).concerning the effect of weight reduction, foods generating weight reduced ranging from 3% towards 9% of one's weight stood to some extent relating to blood pressure declines around 3 mm Hg systolic and diastolic were observed. This methodical assessment was with inadequate command to spot changes in illness or death results (Mulrow et al., 2004).

2.4 Physical Activity

Goldberg and Eckstein (2012)defined physical activity as some human bodily movement which result in outflow of energy. This bodily movement can be characterized into occupational (bodily movement at work place), leisureliness (non-professional bodily movement), exercise (planned bodily movement for a specific purpose) and vigorous lifestyle (eg housekeeping). Bodily movements are frequently referred to as containing three scopes: period (eg 60 minutes), rate (eg for every week) and greatness (eg. percentage of energy outflow)(Montoye, 1996).Consistent physical activity has both protective and healing effects on several prolonged disorders like stroke (Schulz et al., 2004).

2.4.1 Physical Activity as a Separate Risk Factor

To determine the consequence of activities that are physical on stroke 10 studies which consisted controlled vital known risk factors were monitored. Detailed components of the research established an opposite relationship concerning possibility of a coronary incident and physical activity (Tanasescu et al., 2002).

The extent of operative stretched commencing unimportant associations on behalf of exact categories of movement (e.g. energetic go back and forth; threat ratio=1.08, 95% CI 0.95 to 1.23) Hu et al. (2004) to very important links (e.g. males that ran aimed at 60 minutes or extra for each week ensured a 42% possibility decline, RR 0.58, 95% CI 0.44 to 0.77) liken to that of males who decided not run (p<.001) (Tanasescu et al., 2002). Multivariate odds ratio of 0.51 (95% CI 0.29 to 0.90) involving two or more variable quantities were identified by some well-directed case control trials after matching small state of professional physical motion against great states (Altieri et al., 2004).

2.5 Smoking as a Risk Factor

Tobacco smoking is a major cause of stroke according to Doll et al. (2004).In Lemogoum et al. (2005a) work, they stated that the probability of developing stroke is doubled by cigarette smoking. Smoking end cuts these risks considerably, even though the lessening is dependent on the duration of ending (Ockene and NH, 1997);Wannamethee et al. (1995).Nurses Health Study specified that Smokers stood at approximately 1.9 times the possibility of developing stroke matched with persons who under no circumstances smoked Colditz et al. (1993).

Two closely monitored studies showed that persons who do not smoke but unprotected to smoke emitting from cigarette are associated with greater than before possibility of serious coronary conditions of 51% (OR = 1.51, 95% CI 1.21 to 2.99) matched by way of persons who do not smoke and with no exposure of smoke (Pitsavos et al., 2002);(2002b).

2.6 High Blood Pressure (Hypertension) as Risk Factor

Risk factor for stroke that remains utmost vital manageable is hypertension (Gillum, 1999). According to the world health information 2002, high blood pressure famous as hypertension exists as a primary grounds of preventable demise and disease among all world regions. Yet, hypertension consciousness, handling and control changes as low as 34.0%, 28.0% and 6.2%, respectively have been associated with Ashanti region of Ghana. (Agyemang, 2006). Existing literature has made known that modest, effective control of high blood pressure only can decrease stroke incidence by as much as 70% (Zumo, 2006).

2.7 Diabetes and Stroke

Diabetes mellitus plus weakened ability to tolerate glucose are now rampantly associated with both industrialized and unindustrialized nations (Lemogoum et al., 2005b). Diabetes incidence among the urban areas in Ghana is about 6.4%, whereas mature men mortality related to diabetes is approximately 34 per 100,000 in Tanzania. Stroke is associated with diabetes with an estimated relative risk of 1.5 to 3.0. Lemogoum et al. (2005b). Patients with diabetes generally have hypertension and hyperlipidemia which further escalate their risk. Reddy and Yusuf (1998). Even though diabetes is remediable, the occurrence of the disease still intensify one's risk of stroke. (Reddy and Yusuf, 1998);(Zumo, 2006).

2.8 The LASSO

A great extent of determination has been keen on the growth of penalized regression techniques aimed at concurrently variable selection and coefficient estimation. In using enormous number of explanatory variables it is very so often to desire the selection of a reduced subgroup which will not solitary model as good by means of the complete variable group, on the other hand also holds the most significant of explanatory variables. Such apprehensions have directed to important growth of LSR regression procedures through several constraints to determine significant predictors and obtaining greater estimate correctness in linear regression (Minjung et al., 2010).Forecasts on or after prognostic models may perhaps be used for numerous reasons in medicine, comprising the identification of a particular illness and therapeutic conclusion, choice of patients for randomized clinical trials, and updating patients and their families (refer e.g. Harrell et al. (1996)).

The utmost simple statistical technique aimed at what is named supervised learning relays a response variable Y to a linear explanatory variables $x_1, x_2, ..., x_p$.

$$Y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \varepsilon \tag{2.1}$$

Then ε is the error term that gives as an indication that even if $x_1, x_2, ..., x_p$ is known it does not typically convey to us precisely what Y will be.The right of equation(2.1)minus ε is frequently mention as the predicted outcome or explanatory variable. These remain stated as linear regression models. If there exist records on the outcome or response variable y_i and the explanatory variables x_{ij} for every one of two or more categories within N persons or situations, the least squares procedure selects a model by means of minimizing the sum of squared errors concerning the outcome and the predicted outcome, over coefficients from the regression model $\beta_0, ..., \beta_p$..

Linear regression (LR) is among the known and to a greatest extent useful tools for data scrutiny. LR makes available an easily understood by now or then influential technique used in fitting the result of a group of factors (distinctive attribute) happening to an outcome variable. Per a considerable or relatively great factor size , one ought to contain all the predictors in the model. Selection of variable remains an important and demanding task associated with regression : selecting the greatest explanatory predictors to be included in the model. Traditional variable selection approaches examine over and done with all combinations of predictors hence take excessively long to calculate once the number of factors is approximately larger than 30; detailed in Hastie et al. (2001).

One method that enables LR to largely solve problems with lots of predictors is the penalized regression technique. In penalized regression the lasso take absolute value of regression coefficients or l_1 as penalties. Specifically , it makes available an influential method aimed at achieving variable selection with enormous number of explanatory variables. Fewer non-zero coefficients is produced for the regression coefficients as a result of sparse solution. Sparsity remains vital for predictive accuracy, and for interpretation of the concluding model.Assuming we have a linear regression with explanatory variables x_{ij}, response values y_i for $i = 1, .., N$ and $j = 1, .., p$, the lasso answers the l_1 - penalized regression hence minimizes

$$\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=i}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=i}^{p}|\beta_j| \tag{2.2}$$

intended for the not known parameters $\beta_0, ..., \beta_p$. The penalty function is indicated as the additional term; the complexity of the model and its fit is balanced by it.The shrinkage parameter λ manages the tradeoff for coefficients. The technique known to statisticians as Cross validation is one of its kind used in determining the value of λ. We will obtain a more sparse vector for the concluding $\hat{\beta}$ depending on how large λ is defined(Tibshirani, 2011).

Minimizing the sum of squares with the constraint below is the same as the lasso problem (2.2).

$$\lambda\sum_{j=i}^{p}|\beta_j| \leq k.$$

There exist a bound parameter k producing similar result for every λ in the lasso problem (2.2). One must not forget that if $\lambda = 0$ or equally enough large values of k produce result identical to the traditional least squares results. Ridge regression and Lasso regression are alike, which takes the constraint

$$\lambda \sum_{j=i}^{p} \beta_j^2 \leq k.$$

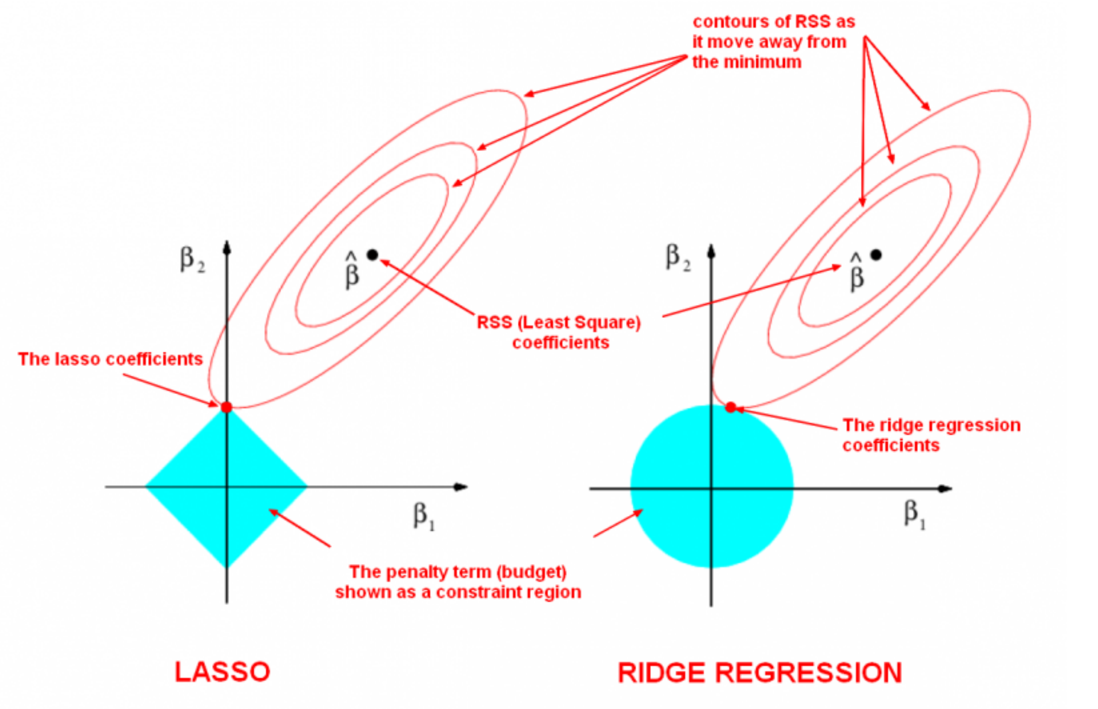


Figure 2.1: Geometric view of LASSO and Ridge

Figure 2.1 displays the error contours and penalized function for both lasso and ridge regression. The dense gray parts are the penalized regions $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s^2$ correspondingly, whereas the ellipses remain the contours of the least squares error function, aligned at the full least squares estimates $\hat{\beta}$.The high-pitched corners of the constraint area meant for the lasso regression produce sparse solutions.Again figure 2.1 provides geometric outlook of the lasso as well as ridge regression.(Tibshirani, 1996)

2.9 Sampled LASSO Background

In latest years at hand is numerous works for lasso regression and l_1 penalization. A list of examples are in table 2.1 modified from Tibshirani (2011). According to Tibshirani his resourceful lasso paper was inspired by knowledge from Leo Breiman's titled the garotte Breiman (1995). The garotte takes $c_1, c_2, ..., c_p$ headed for minimizing

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} c_j x_{ij} \hat{\beta}_j \right)^2 + \lambda \sum_{j=1}^{p} c_j \tag{2.3}$$

where $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_p$ remain the traditional least squares estimates and $c_j \geq= 0$ intended for every $j \in [1, ..., p]$. As a result Leo's idea was headed for scaling the least squares estimates by way of non-negative constants,some of which may perhaps be zero.

The absolute value constraint used by Tibshirani in regression may not be a fresh knowledge entirely. In Chen et al. (1998) "Basis Pursuit "was suggested ,on behalf of signal processing they used absolute value constraints.Before, Frank and Friedman (1993) had (in a nutshell) deliberated the "bridge"estimate,which recommended an assembly of penalties of the prescription $\sum |\beta|^q$ for some q

lasso regression did not get great concern although it was put out in 1996 due to inadequate computational command that was accessible to the mediocre statistician may be the reason, as well as also the fairly small scope of datasets which may not be the situation nowadays. Nowadays, not in statistics only, but likewise in computer science, machine learning and engineering the lasso and l_1 penalty has become a vital area. The drive intended for lasso was interpretability at first:it remains another to subset regression intended for attaining a sparse (or parsimonious) model. In the previous twenty years some unanticipated gains of convex l_1-penalized methods developed: statistical and computational competence.

Concerning the statistical facet , at hand is remarkable effort on the

Table 2.1: Some studies aimed at generalizations of the lasso (l_1) penalty examples

Technique	Researcher(s)
Dantzig selector	Candes and Tao (2007)
Compressive sensing	Donoho (2004),Candes (2006)
Elastic net	Zou and Hastie (2005)
Fused lasso	Tibshirani et al. (2005)
Generalized lasso	Tibshirani and Taylor (2011)
Hierarchical interaction models	Bien et al. (2013)
Multivariate methods	Joliffe et al. (2003)

mathematical characteristics of the lasso,inspecting its capability to improve a true principal (sparse) model and to yield a model by means of insignificant prediction error. Hastie et al. (2001) devised the familiar "Bet on Sparsity"principle in unfolding this work. In identical base L_1 receive that exact sparse. If and only if the assumption remains correct, then and there that parameters will be possibly estimated using l_1 penalties proficiently.If there is a situation where the assumption does not remain correctly compact, then on no account can a technique be able to get better model devoid of a large data size for each parameter.

The convexity difficulty and sparsity of the concluding model which is highly associated with its computation can be used as an improvement. Parameters estimates that are Zero(0) in the response remain controlled by means of minimal cost in the examination. Dominant and accessible methods intended for convex optimization can be useful to the difficulty, letting the solution of very great difficulties. One predominantly effective method that exists is coordinate descent Fu (1998) and Friedman et al. (2010), a modest one-at-a-time process that is suitable for independent lasso constraint.This process is modest and flexible, appropriate for numerous penalized generalized linear models, comprising Logistic, Poisson and Cox's regression for survival data.Coordinate descent is executed in the penalized and glmnet packages just to mention a few in the R statistical software.

2.10 Generalized Linear Models: Logistic Regression

Given a group of variable or response the logistic regression goal is to model the posterior probability of these group of variables Y, $p(Y = d|x)$ by means of transforming the inputs linear combination.For a D-classes problem, D-1 logit functions remain definite as

$$log\left[\frac{p(Y = d|x)}{p(Y = D|x)}\right] = \beta_0^{(d)} + x^t\beta_1^{(d)}, d \in \{1, 2, ..., D - 1\} \tag{2.4}$$

where $(\beta_0^{(d)}, \beta_1^{(d)})$ remain the logistic linear regression estimates used for the d-th class value. The denominator remains set of the D-th class value, on the other hand it may possibly be any. this produces

$$p(Y = d|x) = \frac{\exp(\beta_0^{(d)} + x^t\beta_1^{(d)})}{1 + \sum\limits_{i=1}^{D-1} \exp(\beta_0^{(d)} + x^t\beta_1^{(d)})}, d \in \{1, 2, ..., D - 1\}, \tag{2.5}$$

and therefore $p(Y = D|x) = 1 - \sum\limits_{d=1}^{D-1} p(Y = d|x)$. The logistic model, containing the slope or intercepts, has (p + 1)(D -1)parameters generally, and the maximum likelihood result can be establish by way of the iteratively re-weighted least squares procedure(IRLS), obtained as a result of Newton's method.(Vidaurre et al., 2013)

Tibshirani (1996) expressed the binary grouping problem as the minimization of the l_1-penalized negative log-likelihood function. Equation 2.6 shows how he formulated the function

$$-\sum_{i=1}^{N}[y_i x_i^t\beta - log(1+\exp(x_i^t\beta)) + \lambda||\beta||_1], \tag{2.6}$$

Where the input xi contains the constant span 1 to fit in the intercept. This problem is resolved by way of relating the original lasso procedure for each phase of the IRLS process.

An important landmark aimed at logistic regression application towards data which are highly dimensional stands the latest development of effective approaches for calculating the whole regularization trail of GLMs with convex loss functions. A well-organized regularization path-following procedure for GLMs on the bases of predictor-corrector approaches of convex optimization is specified in Park and Hastie (2007). On the other hand, effective coordinate descent technique for computing the GLM regression coefficients estimates based on a grid of λ parameter values for elastic net penalties have been propose byFriedman et al. (2010). A broad list of advanced processes meant for the sparse logistic regression problem are found in Shi et al. (2010). Again, they put forward a procedure including two phases: a fast iterative shrinkage part and a precise interior point part.

According to Zou (2006) adaptive lasso knowledge remains suitable for logistic regression in precise and generalized linear models entirely, so that the resultant models appreciate the enriched theoretical possessions of the adaptive penalty under confident minor conditions .

2.11 Standard Errors

Standard errors can without difficulty be computed for regression coefficient point estimates. In estimation by penalized technique the variance of the estimate is reduced, whereas bias is introduced considerably making the bias of each estimate a significant factor of its mean square error. Unfortunately, to computing satisfactory and exact bias estimate is difficult and almost impossible for most penalized regression applications thus LASSO. Any bootstrap-based computation can merely define the estimate variance, usually in conditions where penalized estimates are used. There is no consistent unbiased estimates obtainable in penalized regression application in the direction of computing reliable bias for estimates.(Goeman, 2014).

Chapter 3

Methodology

The data analysis was done in two parts: the first part was made up of purely descriptive analyses using Statistical Product and Service Solution (SPSS: IBM version 20). The second part was also made up of inferential statistics which focused on LASSO Logistic regression; parameter estimation odds ratios derivation, predicted probabilities and cross validation of model implemented by R statistical software for windows version.3.1.1. The agreement of proposed methods of weight was done by the use of Bland and Altman test with R software and Cronbach's alpha was carried out by statistical Product and Service solution (SPSS: version 20) to measure the reliability of weight application methods.

3.1 Data

This work used secondary data centred on World Health Organization's (WHO) multi country Study on global Ageing and adult health (SAGE) Wave 1 concerning Ghana, whose discoveries on domestic features about the populace is detailed formerly in Kowal et al. (2012). The 2007/08 SAGE Ghana Wave 1 builds on the 2003/04 World Health Survey (WHS), famously called Sage Wave 0. Sage Wave 1 data gathering included follow-up respondents in use as of Wave 0, and additional fresh respondents making a total of 5573 to increase the cohort scope for upcoming waves.

3.2 Variable Measured

The variables used in this work include classification variables: location (urban or rural), education level, marital status, age, gender and ethnicity. Chronic disease collected were arthritis, angina, asthma, diabetes and hypertension. Lifestyle was also measured: which included smoking alcohol intake, diets record and physical activities. Body Mass Index (BMI) was subsequently calculated by the height and the weight. The absence or presence of stroke was used as binary dependent variable.

For some covariates, the missing value mechanism for the unobserved data used was assumed to be missing at Random (MAR); as is inevitable with spontaneous reporting some information was missing. For the purposes of the analysis presented in this work the final data set for analysis was obtained by firstly removing all rows that contained missing values(be it the covariates), in order to simplify the modelling process.

3.3 Sampling Design

In Sage Wave 1 Ghana deployed stratification and multi stage cluster sampling technique. To obtain sample from Ghana's target population,stratification was undertaken by means of administrative region (Ashanti, Brong Ahafo, Central, Eastern, Greater Accra, Northern, Upper East, Upper West, Volta, Western) and the kind of vicinity (urban/rural), resulting in 20 strata nationally representative. From the strata according to size 10-15 Enumeration Areas were selected.

3.4 The Generalized Linear LASSO Model

From GLM system of equations we can symbolise the outcome ,the beta values and the errors as a vector and the set of explanatory variables as a matrix:

$$Y = X\beta + \varepsilon$$

After a specified data **Y** and the design matrix **X**, the GLM modelling technique has to determine a group of beta values clarifying the data as appropriate as possible:

$$\hat{y} = X\beta$$

A blameless fit would be attained with beta values prominent to predicted values which are as close as possible to the measure Y. When the system of equation is rearranged , it is obvious that a blameless estimate of the data implies small error values:

$$SSE = \sum \left(y - \sum_{j=1}^{p} y\beta_j x_{ij} \right)^2$$

from the equation above we obtain equation 3.1 below:

$$F(\beta, X, Y\lambda) = \sum \left(y - \sum_{j=1}^{p} y\beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{3.1}$$

we can display F as a matrix from 3.1

$$= (Y - X\beta)^T (Y - X\beta) + \lambda \mid \beta \mid I_p \tag{3.2}$$

We can obtain best estimate of β which we can denote by $\hat{\beta}$ when minimizing F. $I_p = p \times p$ identity matrix, $\mid \beta \mid = diag(\mid \beta_j \mid)$ be the diagonal matrix with jth diagonal element, X is the design matrix.

$$|\beta| = \left\{ \begin{matrix} |\beta_1| & 0 & \cdots & 0 \\ 0 & |\beta_2| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & |\beta_p| \end{matrix} \right\} = diag(|\beta|)$$

$$|\beta_j|^{-1} = \left\{\begin{matrix} |\beta_1|^{-1} & 0 & \cdots & 0 \\ 0 & |\beta_2|^{-}1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & |\beta_p|^{-1} \end{matrix}\right\} = diag(|\beta_j|^{-1})$$

we note that equation 3.1 can be express as

$$F(\beta, X, Y, \lambda) = Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta + \lambda \mid \beta \mid \tag{3.3}$$

partial derivative of 3.3 with respect to β is taken and yielded

$$\frac{\partial F(\beta, X, Y, \lambda)}{\partial \beta} = -2Y^T X + 2(X^T X)\beta + \lambda sign(\beta) \tag{3.4}$$

where $sign(\beta) = \left\{\begin{matrix} sign(\beta_1) \\ \vdots \\ sign(\beta_p) \end{matrix}\right\}$

we then set

$$\frac{\partial F(\beta, X, Y, \lambda)}{\partial \beta} = 0 \tag{3.5}$$

and solve for $\hat{\beta}$ we assume that X is orthonormal,$X^T X = I$ for simplicity. $\hat{\beta}_j^{OLS}$ denotes the OLS estimate,for multivariate data

$$\hat{\beta}_j^{OLS} = (X^T X)^{-1} X^T Y = X^T Y$$

.

To find an answer to equation 3.5 we acquired the following result

$$\hat{\beta}_l^{LASSO} = \hat{\beta}_l^{OLS} - \frac{\lambda}{2} sing(\hat{\beta}_l^{OLS}) \tag{3.6}$$

Lemma

if $y + \frac{\lambda}{2} sign(y)$,then x and y share the same sign.

$$y = x - \tfrac{\lambda}{2} sign \Leftrightarrow x = y + \tfrac{\lambda}{2} sign$$

if y is positive , $y + \frac{\lambda}{2} sign$ must be positive then x is positive. on the contrary ,if y is negative then x is negative. Thus x and y share the same sign.
In accordance to lemma above and equation 3.6 $sign(\hat{\beta}_j^{LASSO}) = sign(\hat{\beta}_j^{OLS})$ then equation 3.6 becomes

$$\begin{aligned}
\hat{\beta}_l^{LASSO} &= \hat{\beta}_l^{OLS} - \frac{\lambda}{2} sing(\hat{\beta}_l^{OLS}) \\
&= \begin{cases} \beta_j^{OLS} - \frac{\lambda}{2} & \text{if } \beta_j^{OLS} \geq 0 \\ \beta_j^{OLS} + \frac{\lambda}{2} & \text{if } \beta_j^{OLS} \leq 0 \end{cases} \\
&= \left(\beta_j^{OLS} - \frac{\lambda}{2}\right) I_{\beta_j^{OLS} \geq 0} + \left(\beta_j^{OLS} - \frac{\lambda}{2}\right) I_{\beta_j^{OLS} \leq 0} \qquad (3.7)
\end{aligned}$$

if follows that $\hat{\beta}_j^{LASSO} = sign(\hat{\beta}_j^{OLS})(\hat{\beta}_j^{OLS} - \frac{\lambda}{2})^+$ j=1...p
$+$ Signifies the positive part of expression inside the parenthesis

3.5 Ordinary Logistic Regression(OLR)

Ordinary Logistic Regression (OLR) may be the alternative when dealing with simple bivariate measures (Hauben et al., 2005). In developing a risk score system the approach generally used is the Ordinary Logistic Regression (OLR) that has the form$Y_i = X_i\beta + \varepsilon$ with Y_i Bernoulli random variable that takes on the values either 0 or 1 thus absence or presence of an event; $X = [1, X_{12}, X_{13}, ..., X_{in}]$. The predictor variables $X_1, X_2 ... X_n$ are the observed quantities and can be continuous or indicator/dummy variables reflecting dichotomous risk factors or categories of risk factors. Based on the regression model $\hat{\beta} = (\beta_1, \beta_2, ...\beta_n)$ are the estimates of the regression coefficients; the errors(ε) has mean zero, non-constant variance and therefore needs not follow normal distribution.

Compared to Ordinary Least Square (OLS) parameter estimates, the LASSO) estimates are generally more accurate. Also, some parameters will be shrunk towards zero (0), allowing for better interpretation of the model obtained and identification of those covariates in the model most strongly associated with the outcome(Tibshirani, 1996).

Since the dependent variable Yi has a binary outcome and a vector of risk factors, we can construct an ordinary logistic regression model of Y_i on X for all $i = (1, ...n)$. π_i is defined as the probability that stroke is observed. Then we have a linear function

$$\log(\pi_i) = \left[\pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] = \varphi$$

$$\varphi = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_n X_{ip} = \beta_0 + \sum_{1}^{p} \beta_k X_{ip}$$

If we back transform φ the probability of observing the present of stroke given data X_i and estimated model parameter $\hat{\beta} = (\beta_0 + \beta_1 + ...\beta_n)^T$ this gives

$$\begin{aligned}\pi_i = (Y_i = 1/xi, \beta) = F(\varphi/x_i^T, \beta) \\ &= \frac{\exp(\varphi)}{1 + \exp(\varphi)} \\ &= \frac{1}{1 + \exp(-\varphi)}\end{aligned}$$

3.6 The LASSO Logistic Regression (LLR) Model

The LASSO is a shrinkage and selection method for regression model (Tibshirani, 1996)) originally applied to Ordinary Logistic Regression. For better interpretation of model and identification of those covariates (risk factors) which remain most intensely related to the outcome, the LASSO will shrink some parameters towards zero (0) by putting some constraints on the model parameters with a specified constant as an upper bound to the sum.

To perform LLR with the data obtained, the maximum likelihood approach is commonly used in calculating the coefficients using the equation below

$\pi_i = \left(Y_i = \dfrac{1}{x_i, \beta}\right) = \frac{1}{1+\exp(-\beta^T x_i)}$ for all $i = 1, ...n$
where $x_i = (1, x)^T$ and $\beta = (\beta_0, \beta_1, \beta_2, ...\beta_p)^T$ are vectors of length p+1 thus $= |v| + 1$. Concerning this notation we can recall that
$\pi_i = (Y_i = 1/X_i\beta) = \frac{1}{1+\exp(-\beta^T x)}$ for all $i = 1, ...n$ the joint-likelihood for all i is

$$L(\beta/y_1...Yn) = \mathit{\Pi}_i^n (\pi)^{y_i}(1-\pi)^{1-y_i} \tag{3.8}$$

substitute the value of π in equation 3.8 we obtain

$$= \mathit{\Pi}_i^n \left(\frac{1}{1+\exp(-\beta^T x_i)}\right)^{y_i} + \left(\frac{\exp(-\beta^T x_i)}{1+\exp(-\beta^T x_i)}\right)^{1-y_i} \tag{3.9}$$

hence the joint log-likelihood is obtain as follows

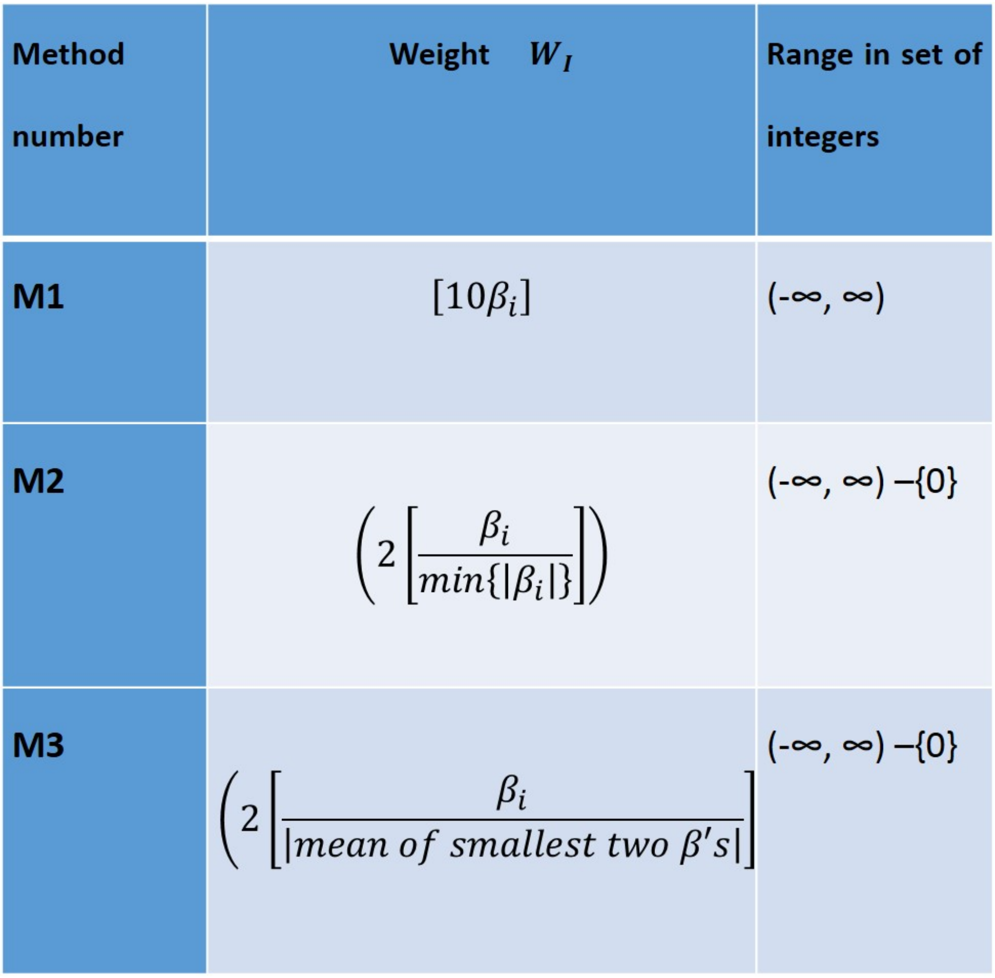
$$\begin{aligned} l(\beta) &= \ln(L(\beta/y_1...Yn)) \\ &= \sum_i^n \left[yi \left(\frac{1}{1+\exp(-\beta^T x_i)}\right) + 1 - y_1 \left(\frac{\exp(-\beta^T x_i)}{1+\exp(-\beta^T x_i)}\right)\right] \quad (3.10) \\ &= -\sum_i^n \left[(1-y_i)\beta^T x_i + \ln(1+\exp(-\beta^T x_i))\right] \quad (3.11) \end{aligned}$$

by adding the LASSO constraint to the joint-likelihood in the 3.11 it yielded the constraint equation below
$l(\beta^{LASSO}) = \sum_i^n \left[(1-y_i)\beta^T x_i + \ln(1+\exp(-\beta^T x_i))\right] + \lambda \sum_{k=1}^p \mid \beta_k \mid$.
The $l(\beta^{LASSO})$ indicates the constrained log-likelihood, where as $\lambda \geq 0$ is the Lagrangian multiplier.

The λ 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^i norm on the ordinary Logistic Regression (OLR) yields the LASSO Logistic Regression (LLR). Some coefficient in the vectorβ become zero depending on how sufficiently λ is specified.

3.7 Methods of Applying Weight W_i

In calculating weights W_i in scoring system the constant w_i is fixed to determine the value of points in the score classification end to end with β coefficients. In risk estimation there exist a wide application of methods; as indicated by the table below



Method number	Weight W_I	Range in set of integers
M1	$[10\beta_i]$	(-∞, ∞)
M2	$\left(2\left[\frac{\beta_i}{min\{\lvert\beta_i\rvert\}}\right]\right)$	(-∞, ∞) –{0}
M3	$\left(2\left[\frac{\beta_i}{\lvert mean\ of\ smallest\ two\ \beta's\rvert}\right]\right.$	(-∞, ∞) –{0}

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

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 β coefficient in the LASSO Logistic Regression model.

This work also identifies two methods M4 and M5 based on LLR model; that is compared to M2. Again, this study has chosen max $|\beta|$ in the LLR model instead of min $|\beta|$ implying that the weight (w_i) for method M4 will be in the range $(-\infty, \infty)$ with $w_i = 100\left(\frac{\beta_i}{max\beta_j}\right)$. In addition, as compared to M3 the study again found the mean of the two absolute largest value of the $\beta's$ to be the denominator Hence for method M5:

$$Wi = \mathbf{100}\left(\frac{\beta_i}{|mean\ of\ largest\ \ two\ \beta's|}\right)$$

then w_i will also be in $(-\infty, \infty)$ range.

3.8 Agreement and Reliability Measure of Scoring Methods

To examine the level of agreement between prevailing approaches and the proposed method the Bland and Altman test was done. The Bland and Altman test estimates the mean difference $\bar{d}$ and standard deviation (s). We assumed that the differences are normally distributed for Gaussian case, then the difference will be situated between $\bar{d}+1.96(s)$ and $\bar{d}-1.96(s)$ thus a 95% confidence interval of differences.(Bland and Altman, 1995).The Bland and Altman test was implemented by the package BlandAltmanleh in R software version 3.1.1

Secondly we also examined the level of constancy through Cronbach's alpha or coefficient alpha. The Cronbach's alpha is a coefficient and it can range from $0-1$ based on the data available; an output of 0 means no consistency whereas 1 means perfect consistency in measurement.Another value that is likely to be obtained in a research is 0.7 meaning that 70% of the variances in the score are reliable variance and therefore 30% is error variance. Bland and Altman (1997). According toBland and Altman (1997) the following rules of thumb: $\geq .9-$ Excellent,$\geq .8-$ Good, $\geq .7-$ Acceptable,$\geq .6-$ Questionable,$\geq .5-$ Poor, and $\leq .5-$ Unacceptable.Cronbach's alpha is express as $\alpha = \dfrac{k\bar{r}}{1+(k-1)\bar{r}}$

where, k = number of indicators. $\bar{r}$ = mean inter- indicators correlation. $k - 1$ = the standardized Cronbach's alpha.

The measuring of reliability of scores was implemented through SPSS for windows ver. 20.

3.9 " Penalized"Package for LASSO Logistic Regression "R".

The LASSO Logistic Regression has been implemented through "R"statistical software for windows ver.3.1.1 using the package penalized. The package penalized was installed in "R". The penalized package is designed for penalized estimation in generalized linear models and also allows an L1 absolute value LASSO penalty (Tibshirani, 1996). Again the penalized package supports other functions that allows the user to find an optimum value of λ (tuning parameter) based on cross-validation and also assesses how parameter estimates change based on the degree of shrinkage or value of λ graphically (Goeman, 2014)

Before fitting the penalized regression model, we test the global null hypothesis of no association between any of the predictor variables and the response by the globaltest package which ties very closely to the penalized package.P value for lasso point estimate were computed using the selectiveinference package.

Chapter 4

Data Analysis and Summary of Findings

Most scientists agree that this non-modifiable things affect the risk of stroke: age, sex, ethnicity (cannot be changed).Also the modifiable risk factors are tobacco smoking, alcohol, weight diet, physical activities, diabetes and hypertension. This chapter gives information on the covariates of stroke modelling using the LLR and the estimation procedure of scores given to a covariate base on the model.

4.1 Exploratory Data Analysis

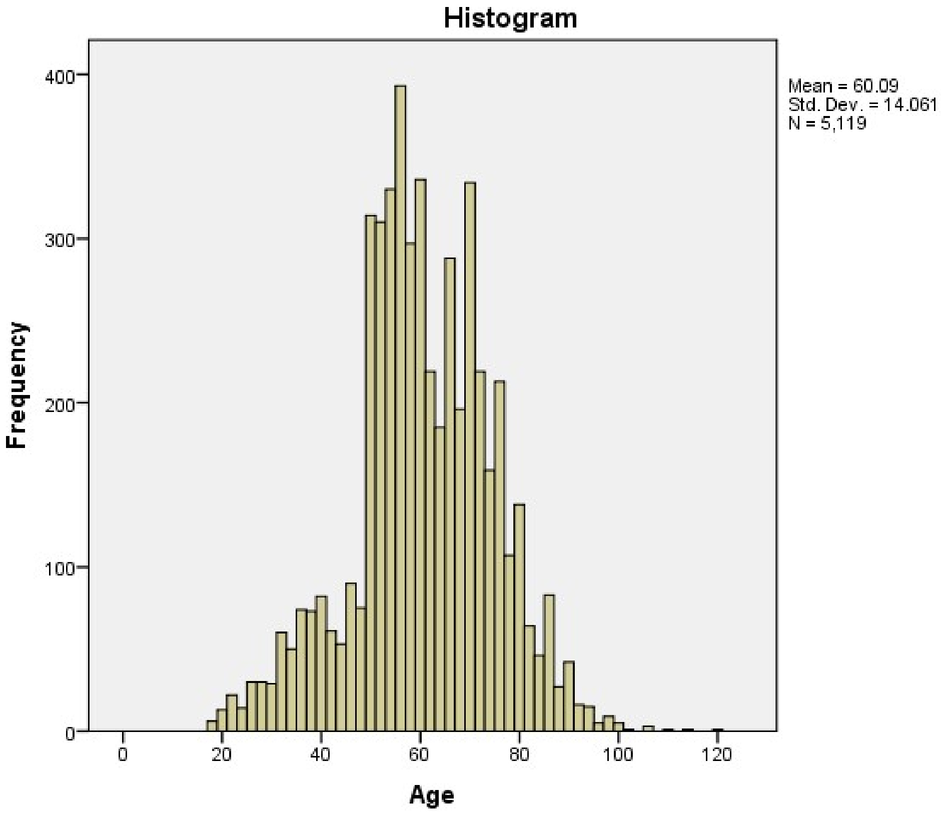


Figure 4.1: Histogram representing the age distribution of all individuals

In exploring the data, tabulation and cross tabulation were embark on in SPSS and graphical designs were produced to review the structure of the data set.Of all 5119 individual contained in the data used for this study, 2714 (53%) existed for male individuals and 2405(47%). Figure (4.1) shows that 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.Appendix A shows the cross tabulation for risk factors and stroke.

4.1.1 Collinearity Statistics

Variable	Tolerance	VIF
Rural/Urban	.876	1.141
Sex	.839	1.192
Mother tongue	.901	1.110
Physical activities	.794	1.259
Tobacco	.819	1.221
Alcohol	.913	1.095
Fruit intake	.937	1.067
Arthritis	.946	1.057
Angina	.979	1.021
Diabetes	.790	1.266
Asthma	.827	1.209
Hypertension	.814	1.228
Chronic lung disease	.851	1.175
WC in centimetres	.994	1.006
Rapid walk	.996	1.004
BMI	.842	1.187
Age	.798	1.254

Based on the collinearity statistics not any of the obtain VIF values is less than 1 or greater than 10 (1 <0r >10) but are all in between 1 to 10 with all tolerance level less than 1, we settle that there is no multicollinearity in the data

4.1.2 Pretesting

Before fitting LLR we tested the global null hypothesis of no association between any of the predictor variables (risk factors) and the response (stroke), against the alternative that there is association between any of the predictor variables (risk factors) and the response (stroke). The test had a p-value $= 4.78e - 15$, a statistic $= 0.655$ with standard deviation (0.0202). Hence the null hypothesis was rejected.

4.2 Optimal Shrinkage (λ) Parameter

K-fold method for cross-validation size was used in obtaining the lambda (λ) 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 λ= 2.988797.

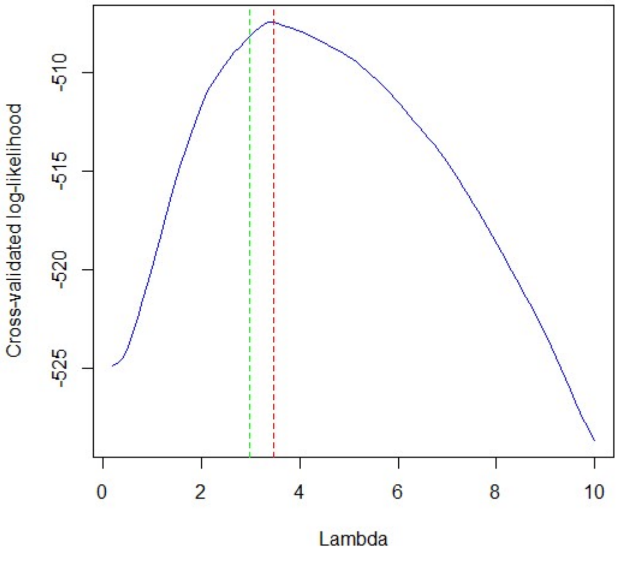


Figure 4.2: Profile of cross-validated log-likelihood for LLR model of Stroke

Figure 4.2 give details about profiling the choice of lambda for LLR model with all risk factors penalized. The optimum value for λ from the "optL1"function within the "penalised"package ($\lambda = 2.988797$) depicted with the green short dashed line. The λ 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.

Following the early examinations into LASSO estimate as against the OLS procedure by Tibshirani (1996) and Hastie et al. (2001),Figure 4.3 displays a line plot of coefficient estimates for the two model mentioned in chapter three. Regarding how changeable the values of coefficient estimates in the LLR model and OLR (Fig.4.3), one detects that an exceedingly large value hypertension, irregular fruit in-take, extreme exercise but shrunk towards zero by the $lamda(\lambda)$ value for the LLR model. This variation was due to the introduction of the LASSO penalty.

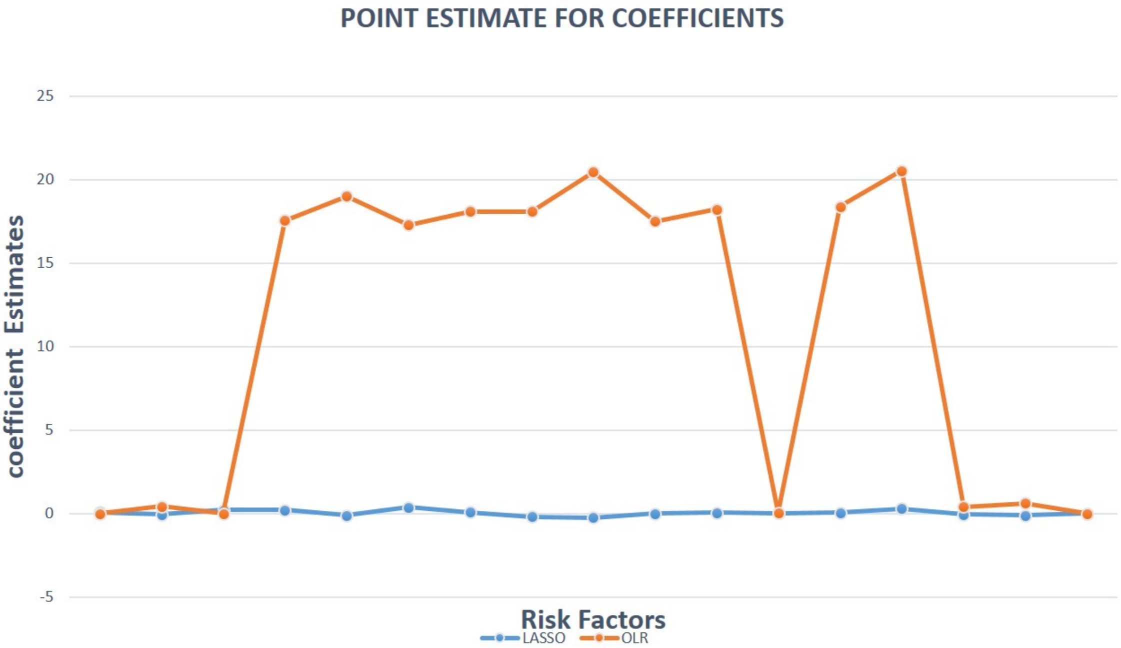


Figure 4.3: OLR and LLR coefficients from models of Stroke

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

In the figure 4.4, 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 λ was then schemed with the "plotpath"command of the "penalised"package. Beginning with the initial the value of λ that marks all the parameters shrunk to zero(0), a pre-defined number of models are created with values of λ not matching another until the last λ is one and the same to that indicated in the set-up of the LLR model.

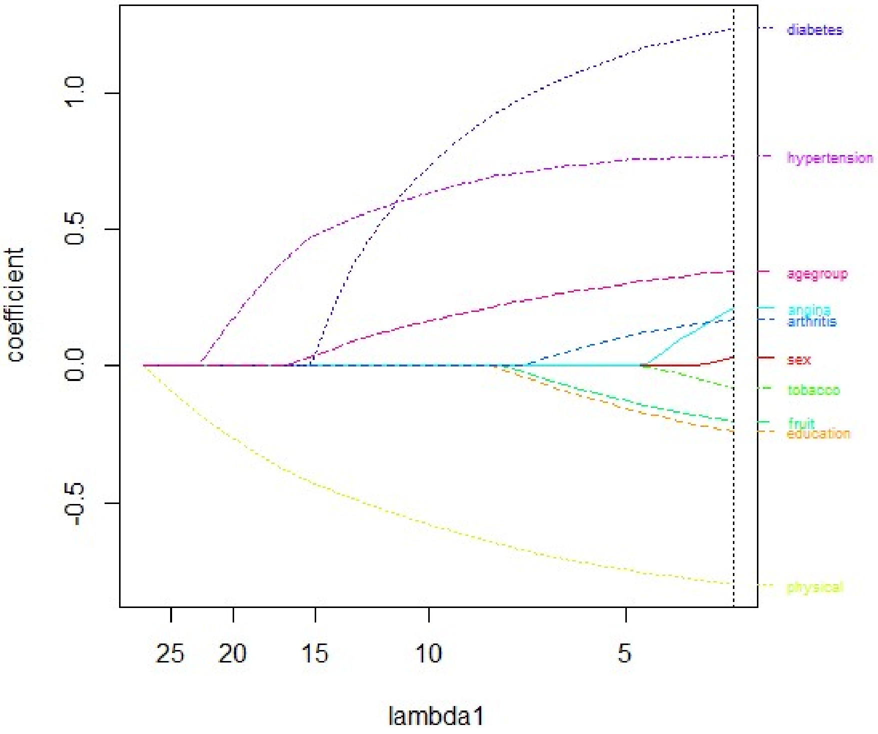


Figure 4.4: Coefficient estimates set against λ values for LASSO Logistic Regression model of Stoke [$\lambda = 3.410935$].

Figure 4.4 was obtained after inputting the fallowing "R"codes using $\lambda = 3.410935$ and 10 steps to display how varying lambda values affect coefficients estimates.

```
>llr = penalized(stroke sex+location+education+physical+alcohol+ethnic+
fruit+angina+arthritis+tobacco+diabetes+hypertension+rapidwalk+bmi+
age + waist, data = bofa, lambda1 = 3.410935, model =“Logistic”, steps=10)
>plotpath(llr,log=”x”)
>abline(v=3.410935, lty=3)
```

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 λ 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.

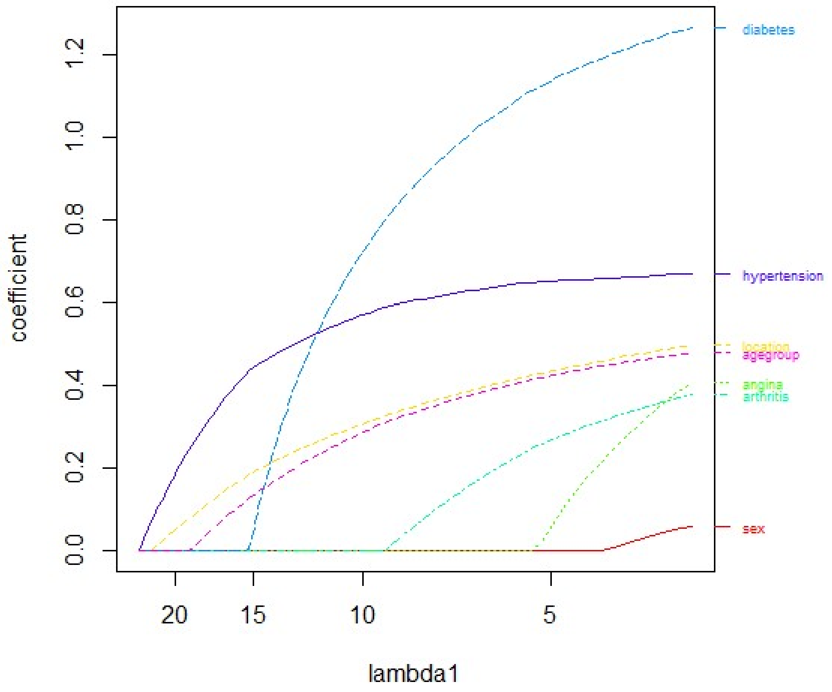


Figure 4.5: Positive Coefficient estimates versus values for LLR model of Stoke [$\lambda = 3.410935$].

For a second time the coefficients were limited to only positive(figure 4.5). Hypertension,dibetes and age were seen to be risk factors intensely associated with the occurrence of stroke, even for values of λ more than 15. Angina and arthritis were identified as an emerging risk factor associated with stroke.

Table 4.1 displays the coefficients labeled 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 or does a mild 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 4.1: **Distribution of risk factors(predictors) and LLR non zero coefficients.**

Risk Factors(Predictors)	LLR Point Estimate	Odds Ratio	P-value
Gender (o=female;1=male)	0.04413259	1.0451209	0.933
Location(0=rural;1=urban)	0.27462354	1.3160351	0.000
Education(0=no formal;1=formal)	-0.05383929	0.9475844	0.130
Physical activities(0=no and mild;1=extreme)	-0.41595175	0.6597121	0.000
Ethnic(o=Akan;1=others)	-0.05333828	0.9480593	0.085
Fruit intake(0=not regular;1=regular)	-0.11438117	0.8919179	0.102
Angina(0=no;1=yes)	0.03743824	1.0381479	0.000
Arthritis(0=no;yes)	0.06544823	1.0676375	0.004
Tobacco (o=used before;1=never)	-0.01857579	0.9815957	0.285
Diabetes(0=no;1=yes)	0.21326979	1.2377185	0.000
Hypertension(0=no;yes)	0.22971266	1.2582384	0.000
Age(0=18-55;1=56+)	0.17412797	1.1902079	0.000

The column in table 4.1 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.

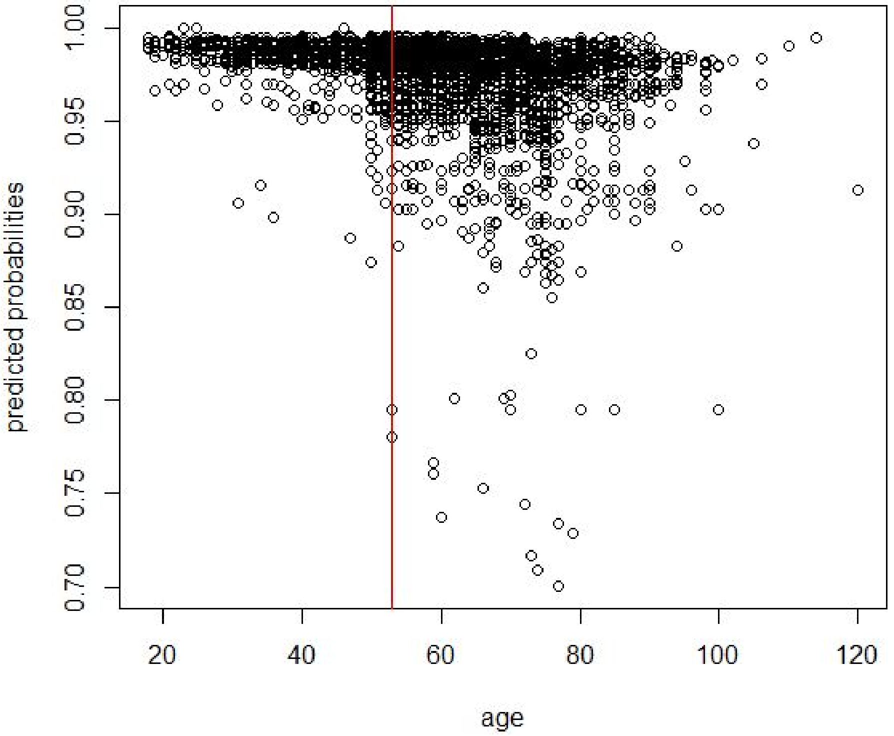


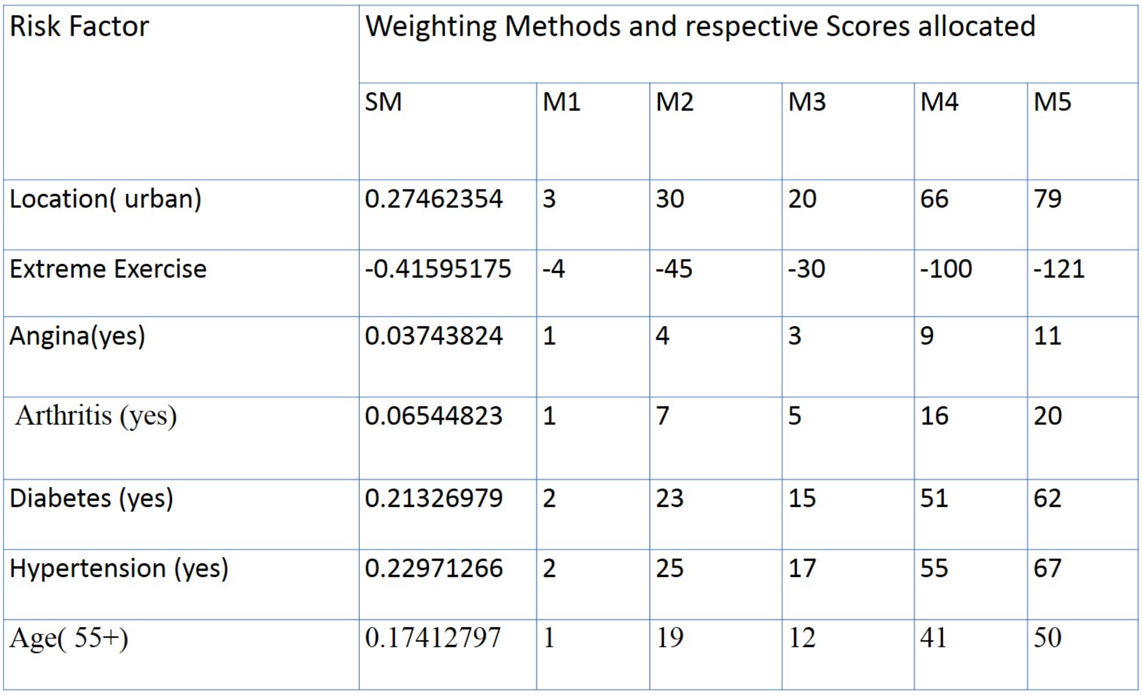
Figure 4.6: Predicted probability against age of an individual

Figure 4.6 shows the predicted probability of individuals 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

4.3.1 Score Allocation

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 of (0.3452876) of the absolute value of the two maximum regression coefficient.

Table 4.2: Points allocated to each risk factor category



Risk Factor	Weighting Methods and respective Scores allocated					
	SM	M1	M2	M3	M4	M5
Location(urban)	0.27462354	3	30	20	66	79
Extreme Exercise	-0.41595175	-4	-45	-30	-100	-121
Angina(yes)	0.03743824	1	4	3	9	11
Arthritis (yes)	0.06544823	1	7	5	16	20
Diabetes (yes)	0.21326979	2	23	15	51	62
Hypertension (yes)	0.22971266	2	25	17	55	67
Age(55+)	0.17412797	1	19	12	41	50

The table above depict the score given by the three existing methods and the extra two proposed by this study detailed in chapter 3.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.

4.4 Test of Agreement for Proposed Methods of Weight Application

Figure 4.7 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).

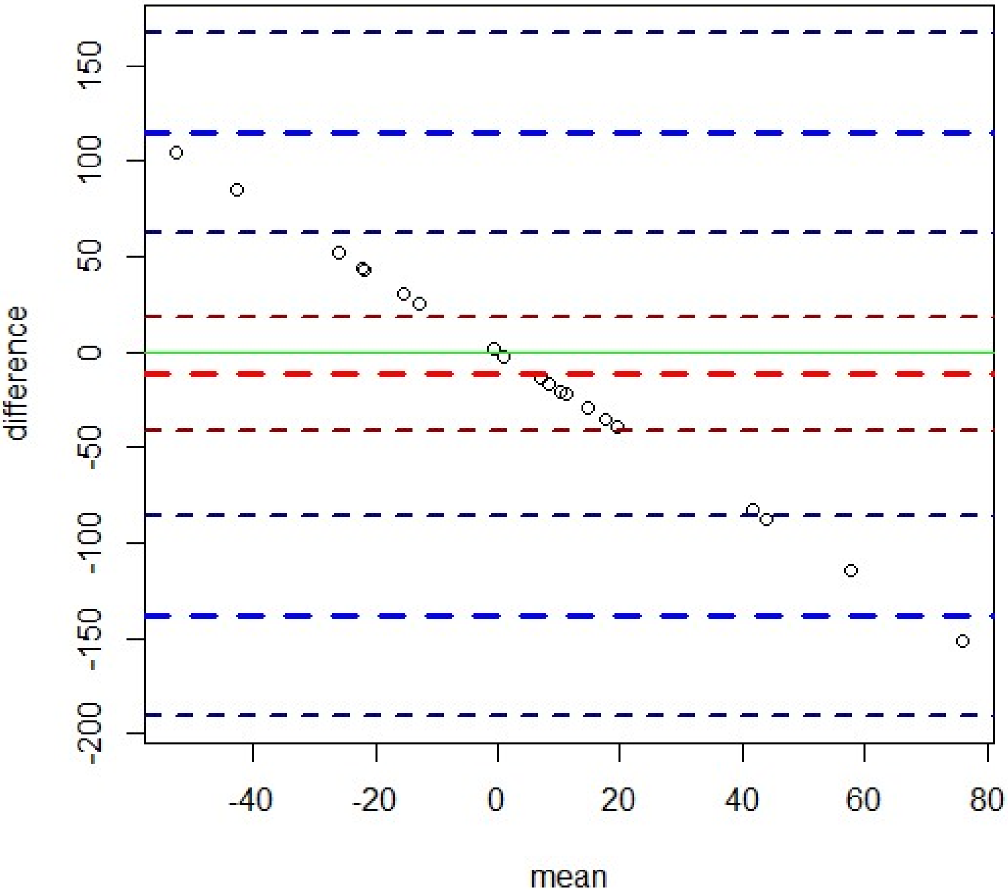


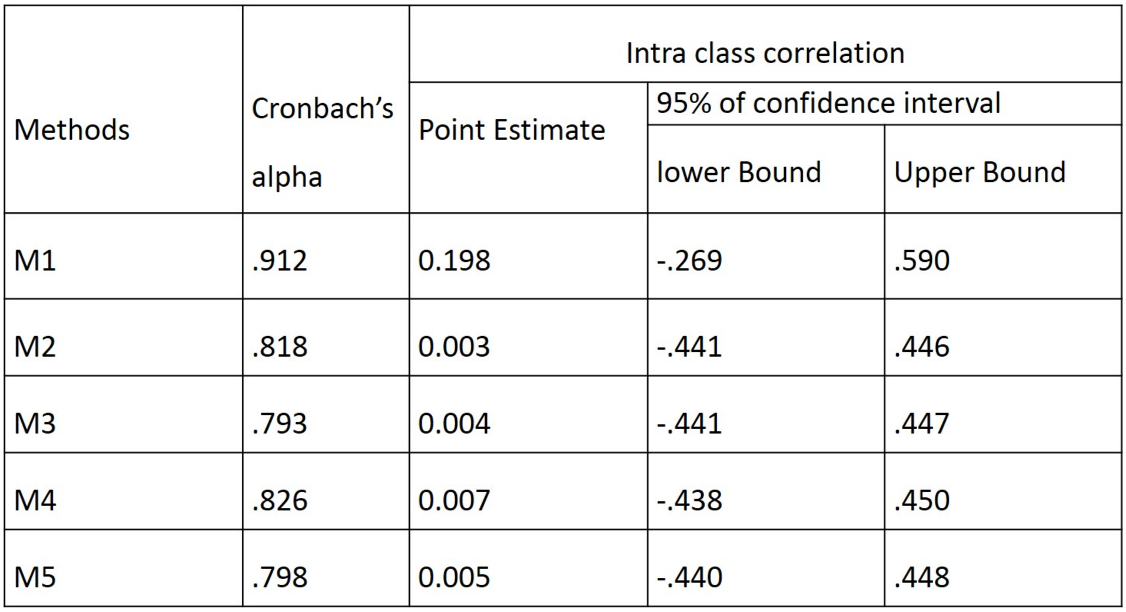
Figure 4.7: Bland and Altman test.

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 M4 and SM. Secondly, the line of equality (0) on the x-axis was also within the error margin (-41.67657, 18.68940) of the mean difference line (-11.49359) for SM and M5. Again the bias was not significant implying a level of agreement for the two methods.

4.5 Reliability Measure for Scoring Methods

The table 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 4.3: Measure of internal consistency among weight application methods



Methods	Cronbach's alpha	Intra class correlation		
		Point Estimate	95% of confidence interval	
			lower Bound	Upper Bound
M1	.912	0.198	-.269	.590
M2	.818	0.003	-.441	.446
M3	.793	0.004	-.441	.447
M4	.826	0.007	-.438	.450
M5	.798	0.005	-.440	.448

Chapter 5

Conclusion and Recommendations

5.1 Introduction

In this chapter we would look at the conclusion and a few other things. From the study we have detailed how we can apply LASSO Logistic Regression (LLR) in modelling the presence or absence of stroke. Recall the two objectives of this study stated in section 1.4.

- To explore the application of LASSO Logistic Regression to select the most informative risk factors of stroke to establish association with stroke.
- To derive a risk scoring method for stroke using statistically acceptable modifications of existing methods.

5.2 Summary

Hypertension was still associated with stroke even at a tuning parameter increased from the optimal value to 25 which gives an indication that hypertension is a factor of stroke. This result from this study is utterly in agreement with the finding that hypertension is a major risk factor (Schulz et al., 2004).Indisputably, the probability plot against age indicated that adults aged 18-53 has a high possibility of developing stroke. This is consistent with a study by Schulz et al. (2004) to determine the risk factors of stroke and Kissela et al. (2012).

The observation from the model tells us that a person who does extreme exercise is less likely to develop stroke relating identical to discovery by Schulz et al. (2004). The study also revealed persons with angina as having high log odds(likely) of developing stroke similar to findings by Krahn et al.

(1995). According to the results arthritis is an emerging risk factor concerning stroke,identical to Ong et al. (2013)

In chapter 4 reliability test for M1-M5 methods with SM specified that M1 has the highest Cronbach's alpha of 0.912 and intra class correlation followed by M4.The Bland and Altman test for agreement indicated that the proposed methods M4 and M5 are all in agreement with SM and the level of bias was not significant for these method compared with SM.

5.3 Conclusion

Data about stroke discovered angina and arthritis as an emerging risk factor of stroke in addition to hypertension,luck physical activities, aging and diabetes. We hypothesise that hypertension, diabetes and lack of physical activities reflect the trend of stroke cases in Ghana.

5.4 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.

REFERENCES

Agyemang, C. (2006). Rural and urban differences in blood pressure and hypertension in ghana, west africa. *Public Health*, 120:522–23.

Aho, K., Harmsen, P., Marquardsen, S. H. J., Smirnov, V. E., and Strasser, T. (1980). Cerebrovascular disease in the community: results of a who collaborative study. *Bull World Health Organisation*, 1980; 58:113–130.

Altieri, A., Tavani, A., Gallus, S., and Vecchia, C. L. (2004). Occupational and leisure time physical activity and the risk of nonfatal acute myocardial infarction in italy. *Ann Epidrmiol*, 14(7):461–6.

Anderson, K. M., Odell, P. M., Wilson, P. W., and Kannel, W. B. (1991). Cardiovascular disease risk profiles. *Am Heart J*, pages 293–298.

Bien, J., Taylor, J., and Tibshirani, R. J. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41:1111–1141.

Bland, J. M. and Altman, D. G. (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet*, pages 346: 1085–1087.

Bland, J. M. and Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, page 314: 572.

Bloom, D. E., Cafiero, E. T., Jane, E. L., Abrahams, G., Bloom, R. L., and Fathima, S. (2011). The global economic burden of non communicable diseases. geneva: World economic forum.

Breiman, L. (1995). Better subset selection using the non-negative garotte. *Technometrics*, 37:738–754.

Candes, E. J. (2006). Compressive sampling. In *Proceedings of the International Congress of Mathematicians, Madrid, Spain.*

Candes, E. J. and Tao, T. (2007). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215.

Capuccio, F. P. (2004). Epidemiological transition, migration and cardiovascular disease. *International Journal of Epidemiology*, 33(2):387–8.

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.*

Chen, S., Donohoand, D. L., and Saunders, M. (1998). Atomic decomposition for basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61.

Colditz, I. K. G. A., Stampfer, M. J., Willett, W. C., Manson, J. E., Rosner, B., and et al (1993). Smoking cessation in relation to total mortality rates in women. a prospective cohort study. *Ann Intern Med.*, 119(10):992–1000.

DeSA, U. N. (2013). World population prospects: the 2012 revision. population division of the department of economic and social affairs of the united nations secretariat, new york.

Doll, R., Peto, R., Boreham, J., and Sutherland, I. (2004). Mortality in relation to smoking: 50 years observations on male british doctors. *BMJ*, 328(7455):1519–27.

Donoho, D. L. (2004). Compressed sensing. Technical report, Statistics Department,Stanford University, Stanford, CA.

Feigin, V. L. (2007). Stroke in developing countries: can the epidemic be stopped and outcomes improved. *Lancet Neurol*, 6:94–97.

Friedman, J., Hastie, T., and Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:Article 1.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416.

Gillum, R. F. (1999). Risk factors for stroke in blacks: A critical review. *American Journal of Epidemiology*, 150(12):1266–74.

Goeman, J. J. (2014). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, pages 52(1), 70–84.

Goldberg, N. and Eckstein, J. (2012). Sparse weighted voting classifier selection and its linear programming relaxations. *Information Processing Letters*, 112:481–486.

Harrell, F. E., Lee, J. K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning: Data mining, inference and prediction. *Springer, Canada.*, 2nd Edition,.

Hauben, M., Madigan, D., Gerrits, C. M., Walsh, L., and van Puijenbroek, E. P. (2005). The role of data mining in pharmacovigilance. *Expert Opinion on Drug Safety Data*, 4(5):pp. 929–948.

Hooper, L., Summerbell, C. D., Higgins, J. P. T., Thompson, R. L., Clements, G., Capps, N., and et al. (2004). Reduced or modified dietary fat for preventing cardiovascular disease (cochrane review). in: The cochrane library, issue 2,chichester; john wiley.

Hu, N. C. B. G., Lakka, T. A., Pekkarinen, H., Nissinen, A., and Tuomilehto (2004). Low physical activity as a predictor for total and cardiovascular disease mortality in middle-aged men and women in finland. eur heart j. 2004;25(24):2204-11. *Eur Haert J*, 25(24):2204–11.

Jackson, R., Lawes, C. M., and Bennett, D. A. (2005). Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *Lancet*, 365:434 – 441.

Joliffe, I. T., Trendafiov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531 –547.

Kissela, B. M., Khoury, J. C., Alwell, K., Moomaw, C. J., Woo, D., Adeoye, O., Flaherty, M. L., Khatri, P., Ferioli, S., Rosa, L. A., R, F. D. L., and Broderick, J. P. (2012). Age at stroke temporal trends in stroke incidence in a large, biracial population. *Neurology*, 79(17),:pp.1781–1787.

Komarek, P. and Moore, A. (2003). *Artificial Intelligence and Statistics.*

Kowal, P., Chatterji, S., Naidoo, N., Biritwum, R., Wu, F., Lopez, R., Maximova, T., Arokiasamy, P., Phaswana, N. M., Williams, S., Snodgrass, J., NMinicuci, D'Este, C., Peltzer, K., and Boerma, J. (2012). the sage collaborators: Data resource profile: The world health organization study on global ageing and adult health (sage). *Int J Epidemiol*, pages 1–11.doi:10.1093/ije/dys210.

Krahn, A. D., J, M., Tate, R. B., and Cuddy, F. A. M. T. E. (1995). The natural history of atrial fibrillation: incidence, risk factors, and prognosis in the manitoba follow-up study. *The American journal of medicine*, 98 (5):pp.476–484.

Lemogoum, D., Degaute, J. P., and Bovet, P. (2005a). Stroke prevention, treatment and rehabilitation in sub-saharan africa. *The American Journal of Preventive Medicine.2005*, 29(51):95–101.

Lemogoum, D., Degaute, J. P., and Bovet, P. (2005b). Stroke prevention, treatment and rehabilitation in sub-saharan africa. *The American Journal of Preventive Medicine*, 29(51):98–101.

Mathers, C. D. and Loncar, D. (2005). Updated projections of global mortality and burden of disease, 2002-2030: data sources, methods and results. evidence and information for policy. geneva, world health organization,.

Minicuci, N., Biritwum, R. B., MENSAH, G., Yawson, A. E., Naidoo, N., Chatterji, S., and Kowal, P. (2014). Sociodemographic and socioeconomic patterns of chronic non-communicable disease among the older adult population in ghana. *Global health action*, page 7.

Minjung, K., Gilly, J., Malay, G., and George, C. (2010). Penalized regression, standard errors, and bayesian lassos. *International Society for Bayesian Analysis*, 5, Number 2:pp. 369–412.

Montoye, H. L. (1996). Measuring physical activity and energy expenditure.cham[aogn (il). *Human Kinetics:*.

Mulrow, C., Chiquette, E., Angel, L., Cornell, J., Summerbell, C., Anagnostelis, B., and et al. (2004). Dieting to reduce body weight for controlling hypertension in adults (cochrane review). in: The cochrane library, issue 2, chichester; john wiley.

Murray, C. V., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., Abdalla, S., and Aboyans, V. (2013). Disability - adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990 - 2010 a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):pp.2197–2223.

Ockene, I. S. and NH, N. H. M. (1997). Cigarette smoking, cardiovascular disease, and stroke: a statement for healthcare professionals from the american heart association. american heart association task force on risk reduction. *Circulation.*, 96(9):3243–7.

Ong, K. L., Wuu, B. J., Cheung, B. M., Barter, P. J., and Rye, K. A. (2013). Arthritis: its prevalence, risk factors, and association with cardiovascular diseases in the united states, 1999 to 2008. *Annals of epidemiology*, 23(2),:pp.80–86.

Park, M. Y. and Hastie, T. (2007). l1-regularization path algorithm for generalized linear models. *J. Roy. Stat. Soc. B*, 69(4):659–677.

Pitsavos, C., Panagiotakos, D. B., Chrysohoou, C., Skoumas, J., K, K. T., Stefanadis, C., and et al (2002). Association between exposure to environmental tobacco smoke and the development of acute coronary syndromes: the cardio2000 case-control study. *Tob Control*, 11(3):220–5.

Reddy, K. S. and Yusuf, S. (1998). Emerging epidemic of cardiovascular disease in developing countries. *Circulation*, 97:596–601.

Rose, G. (1981). Strategy of prevention: lessons from cardiovascular disease. *Br Med J (Clin Res Ed)*, 282(6279)::1847–51.

Schulz, U.G.R, Flossmann, E., and Rothwell, P. M. (2004). Heritability of ischemic stroke in relation to age, vascular risk factors, and subtypes of incident stroke in population-based studies. *Stroke*, 35(4):819–824.

Shi, J., Yin, W., Osher, S., and Sajda, P. (2010). A fast hybrid algorithm for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res*, 11(1):713–741.

Stamler, J., Stamler, R., and Neaton, J. D. (1993). Blood pressure, systolic and diastolic, and cardiovascular risks: Us population data. *Arch Intern Med*, 153(5):598–615.

Sudlow, C. L. M. and CP., C. P. W. (1996). Comparing stroke incidence worldwide: what makes studies comparable? *stroke*, 27:550–558.

Tanasescu, M., Leitzmann, M. F., Rimm, E. B., Willett, W. C., Stampfer, M. J., and Hu, F. B. (2002). Exercise type and intensity in relation to coronary heart disease in men. *JAMA*, 288(16):1994–200.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.*, Series B (Methodological):pp. 267–288.

Tibshirani, R. J. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society, Series B*, 73:273–282.

Tibshirani, R. J., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *journal of the Royal Statistical Society, Series B*, 67:91–108.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. In *Technical Report,Stanford University, Stanford, CA.*

Vaitheeswaran, K., Subbiah, R. R. M., and Raman, R. (2014). Risk score estimation of diabetic retinopathy: Statistical alternativesn using multiple logistic regression. *J Biomet Biostat*, 5: 211:doi:10.4172/2155–6180.1000211.

Vestfold, H. S. G. (2003). Influence on lifestyle measures and five-year coronary risk by a comprehensive lifestyle intervention programme in patients with coronary heart disease. *ur J Cardiovasc Prev Rehabil*, 10(6):429–37.

Vidaurre, D., Bielza, C., and p Larrañaga (2013). A survey of l1 regression. *International Statistical Review*, 81(3):361–387.

Wannamethee, S. G., Shaper, A. G., Whincup, P. H., and Walker, M. (1995). Smoking cessation and the risk of stroke in middle-aged men. *JAMA*, 274(2):155–60.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J Amer Statistics Association*, 101(12):1418 – 1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Series B (Statistical Methodology) 67(2),:30–320.

Zumo, A. (2006). What is stroke? how to recognize, prevent and treat stroke. *The Perspective*.

Appendix A

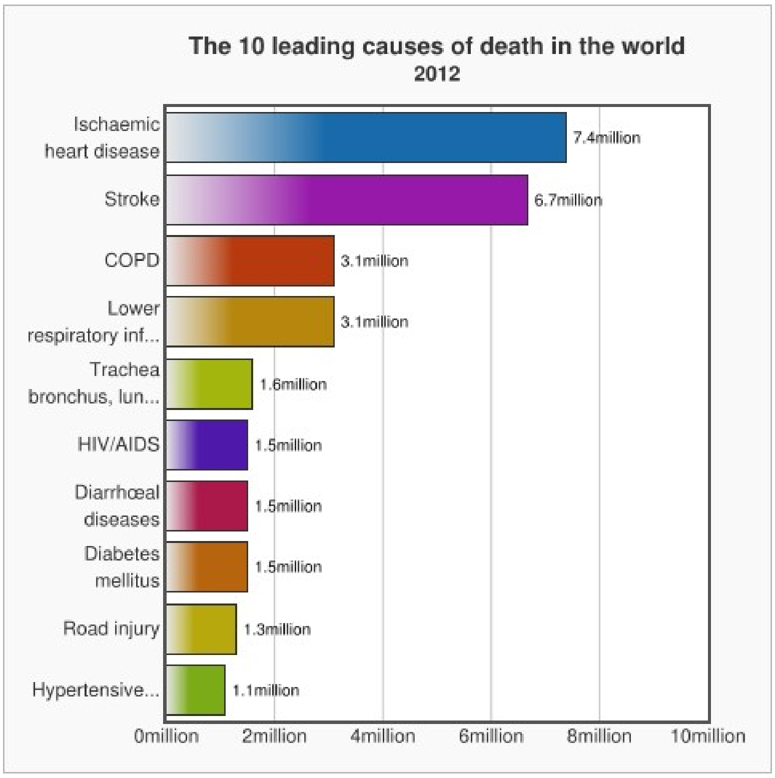
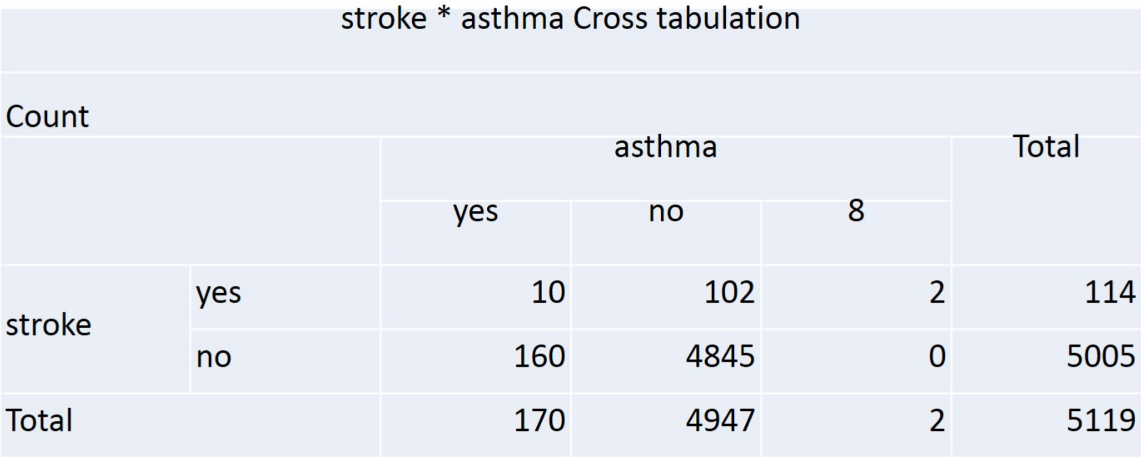


Figure 5.1: The 10 leading causes of death in the world, 2000 and 2012

Cross tabulation on asthma as risk factor of stroke

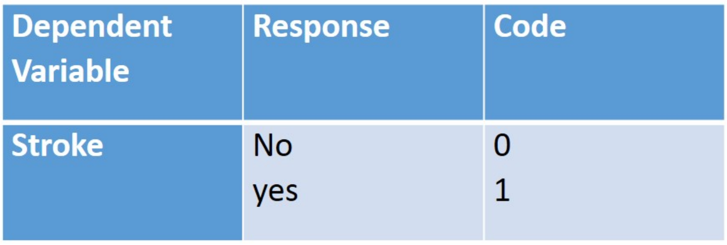


stroke * asthma Cross tabulation

Count

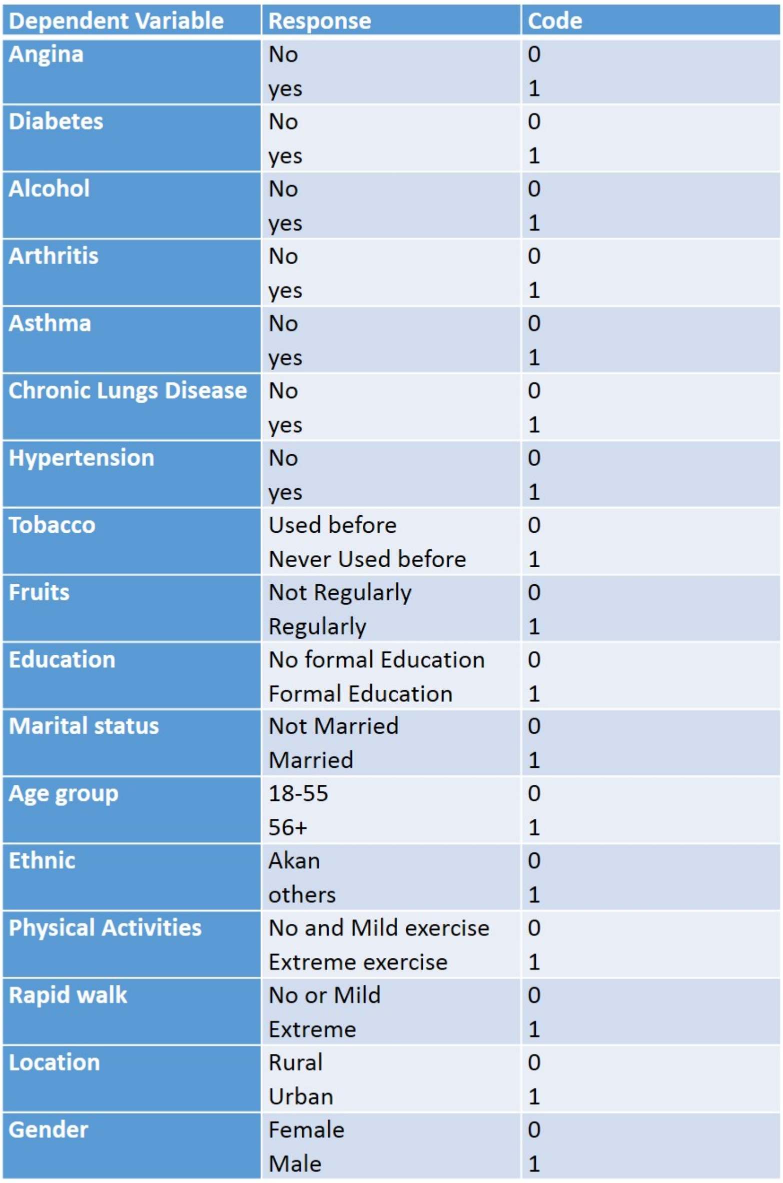
		asthma			Total
		yes	no	8	
stroke	yes	10	102	2	114
	no	160	4845	0	5005
Total		170	4947	2	5119

Coding of response Variables



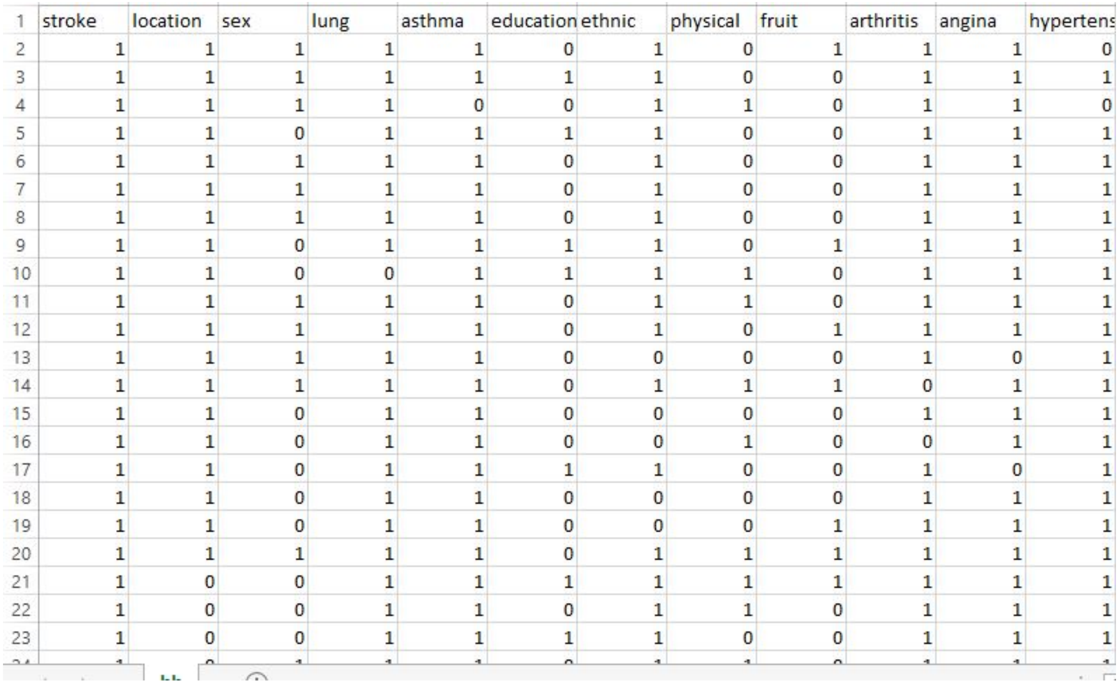
Dependent Variable	Response	Code
Stroke	No yes	0 1

Coding of Explanatory Variables



Dependent Variable	Response	Code
Angina	No yes	0 1
Diabetes	No yes	0 1
Alcohol	No yes	0 1
Arthritis	No yes	0 1
Asthma	No yes	0 1
Chronic Lungs Disease	No yes	0 1
Hypertension	No yes	0 1
Tobacco	Used before Never Used before	0 1
Fruits	Not Regularly Regularly	0 1
Education	No formal Education Formal Education	0 1
Marital status	Not Married Married	0 1
Age group	18-55 56+	0 1
Ethnic	Akan others	0 1
Physical Activities	No and Mild exercise Extreme exercise	0 1
Rapid walk	No or Mild Extreme	0 1
Location	Rural Urban	0 1
Gender	Female Male	0 1

Sample of Data



1	stroke	location	sex	lung	asthma	education	ethnic	physical	fruit	arthritis	angina	hypertens
2	1	1	1	1	1	0	1	0	1	1	1	0
3	1	1	1	1	1	1	1	0	0	1	1	1
4	1	1	1	1	0	0	1	1	0	1	1	0
5	1	1	0	1	1	1	1	0	0	1	1	1
6	1	1	1	1	1	0	1	0	0	1	1	1
7	1	1	1	1	1	0	1	0	0	1	1	1
8	1	1	1	1	1	0	1	0	0	1	1	1
9	1	1	0	1	1	1	1	0	1	1	1	1
10	1	1	0	0	1	1	1	1	0	1	1	1
11	1	1	1	1	1	0	1	1	0	1	1	1
12	1	1	1	1	1	0	1	0	1	1	1	1
13	1	1	1	1	1	0	0	0	0	1	0	1
14	1	1	1	1	1	0	1	1	1	0	1	1
15	1	1	0	1	1	0	0	0	0	1	1	1
16	1	1	0	1	1	0	0	1	0	0	1	1
17	1	1	0	1	1	1	1	0	0	1	0	1
18	1	1	0	1	1	0	0	0	0	1	1	1
19	1	1	0	1	1	0	0	0	1	1	1	1
20	1	1	1	1	1	0	1	1	1	1	1	1
21	1	0	0	1	1	1	1	1	1	1	1	1
22	1	0	0	1	1	0	1	1	0	1	1	1
23	1	0	0	1	1	1	1	0	0	1	1	1

Appendix B

Glossary

BP	Blood Pressure
CVD	Cardiovascular Disease
DALYs	Disability Adjusted Life Years
GLM	Generalized linear model
IRLS	Iteratively Reweighed Least Square Algorithm
LASSO	Least Absolute Shrinkage and Selection Operator
LSR	least Square Regression
LR	linear Regression
LLR	Lasso Logistic Regression
MLE	Maximum Likelihood Estimate
OLS	Ordinary Least Square
OLR	Ordinary Logistic regression
SM	Standard Method

“R”Code for Producing LLR Model

Choice of lambda

```
>library(penalized)
>optimum=optL1(stroke sex+education+marital+physical+tobacco+alcohol
+fruit+angina+diabetes+hypertension+rapidwalk+bmi+agegroup, data=bofa,
minlambda1=0.2, maxlambda1=10, lambda2=0,model=”logistic”,fold=10)
>optimum$lambda
```

gives optimum lambda value by optL1 function lambda1 = 2.988797

```
>profile=profL1(stroke sex+education+marital+physical+tobacco+alcohol+fruit
+angina+diabetes+hypertension+rapidwalk+bmi+agegroup, data=bofa,
minlambda1=0.2, maxlambda1=10, lambda2=0,model=”logistic”,fold=10)
>profile $cvl
```

>profile $lambda

>plot(profile$lambda, profile$cvl, type=”l”, col=”blue”, xlab=”Lambda”, ylab=”Cross-validated log-likelihood”)

>abline(v=2.988797, lty=2,col=”green”)

>abline(v=3.46428 , lty=2,col=”red”)

The code further down is used to fit and evaluate stroke

>fit=penalized(stroke 0+sex+location+education+martal+physical+ alcohol +ethnic+fruit+angina+arthritis+tobacco+diabetes +hypertension+rapidwalk +bmi+agegroup+waist,data=bofa, lambda1=3.410935 ,model=”logistic”)

>coefficients(fit) ⇒ this gives coefficient estimates.

Below allows “plotpath” to be used to see how changing lambda values affect coefficients estimates .

>llr=penalized(stroke sex+education+marital+physical+tobacco+alcohol+fruit +angina+diabetes+hypertension+lung+rapidwalk+bmi+agegroup, data=bofa, lambda1=3.410935 ,model=”logistic”,steps=100)

>plotpath(llr,log=”x”)

Codes for Bland and Altman test

>library (BlandAltmanLeh)

>bland.altman.plot(tsm, tM1,two = 1.96, mode = 1,graph.sys = ”base”, conf.int =0.95, silent = TRUE, sunflower = FALSE,xlab=”Means”,ylab=”Difference”)