Hindawi Publishing Corporation
International Journal of Quality, Statistics, and Reliability
Volume 2008, Article ID 471607, 10 pages
doi:10.1155/2008/471607

Research Article

Sensitivity Analysis to Select the Most Influential Risk Factors in a Logistic Regression Model

Jassim N. Hussain

School of Mathematical Sciences, University Sains Malaysia, 11800 Penang, Malaysia

Correspondence should be addressed to Jassim N. Hussain, j_nassir2000@yahoo.com

Received 1 August 2008; Revised 17 October 2008; Accepted 25 November 2008

Recommended by Myong K. (MK) Jeong

The traditional variable selection methods for survival data depend on iteration procedures, and control of this process assumes tuning parameters that are problematic and time consuming, especially if the models are complex and have a large number of risk factors. In this paper, we propose a new method based on the global sensitivity analysis (GSA) to select the most influential risk factors. This contributes to simplification of the logistic regression model by excluding the irrelevant risk factors, thus eliminating the need to fit and evaluate a large number of models. Data from medical trials are suggested as a way to test the efficiency and capability of this method and as a way to simplify the model. This leads to construction of an appropriate model. The proposed method ranks the risk factors according to their importance.

1. Introduction

Sensitivity analysis (SA) plays a central role in a variety of statistical methodologies, including classification and discrimination, calibration, comparison, and model selection [1]. SA also can be used to determine which subset of input factors (if any) accounts for most of the output variance (and in what percentage); those factors with a small percentage can be fixed to any value within their range [2]. In such usage, the focus is on determination of the important variables to simplification of the model; the original motivation for our research lay in a search for how to best arrive at such a determination. Although SA has been widely used in normal regression models to extract important input variables from a complex model so as to arrive at a reduced model with equivalent predictive power, it has limited use for selection of risk factors despite the presence of a large number of risk factors in survival regression models. The limited use of these methods to select appropriate subsets in survival regression models illustrates the desirability of development of a new method of SA-based variable selection to avoid the drawbacks of traditional methods and also simplify survival regression models by choosing the appropriate subsets of risk factors.

A considerable number of methods of variable selection have been proposed in the literature. The fundamental developments are squarely in the context of normal regression models and particularly in the context of multivariate linear regression models [3]. A comprehensive review of many variable selection methods is represented in [4]. Methods such as forward, backward, and stepwise selection and subset selection (Akaike information criterion (AIC)) and Bayesian information criterion (BIC)) are available; however none of these methods can be recommended for use in either a logistic regression model or in other survival regression models. They give incorrect estimates of the standard errors and P-values. They also can delete variables whose inclusion is critical [3]. In addition, these methods regard all the risk factors of a situation as equal, and they seek to identify the candidate subset of variables sequentially; furthermore, most of these methods focus on the main effects and ignore higher-order effects (interactions of variables).

New methods of variable selection, such as *least absolute shrinkage and selection operator* (LASSO) in [5], and the *smoothly clipped absolute deviation* (SCAD) method in [6], are at the center of attention recently in the field of survival regression models. These methods use the penalized likelihood estimation and the shrinkage regression approaches.

These two approaches differ from traditional methods in their deletion of the nonsignificant covariates in the model by estimating their effects as 0. A nice feature of these methods is that they perform estimation and variable selection simultaneously, but, nevertheless, these methods suffer from some calculation and characteristics problems that are dealt with in more detail in [7, 8].

This study aims to use SA to extend and develop an effective, efficient, and time-saving variable selection method in which the best subsets are identified according to specified criteria without resorting to fitting all the possible subset regression models in the field of survival regression models. The remainder of this study is organized as follows: Section 2 gives the background of building a logistic regression model, and Section 3 deals with the proposed method. The results of implementing this method and logistic regression model are the subject of Section 4, and Section 5 consists of the discussion and conclusions.

2. Background of Constructing a Logistic Regression Model

Often the response variable in clinical data is not a numerical value but a binary one (e.g., alive or dead, diseased or not diseased). When the latter occurs, a binary logistic regression model is an appropriate method to present the relationship between the disease's measurements and its risk factors. It is a form of regression used when the response variable (the disease measurement) is a dichotomy and the risk factors of the disease are of any type [9]. A logistic regression model neither assumes the linearity in the relationship between the risk factors and the response variable, nor does it require normally distributed variables. It also does not assume homoscedasticity, and in general has less stringent requirements than linear regression models. However, it does require that observations are independent and that the independent risk factors are linearly related to the logit of the response variable [10]. However, models involving the association between risk factors and binary response variables are found in various disciplines such as medicine, engineering, and the natural sciences. How do we model the relationship between risk factors and binary response variable? The answer to this question is the subject of the next subsections.

2.1. Constructing a Logistic Regression Model

The first step in modeling binomial data is a transformation of the probability scale from range $(0, 1)$ to $(-\infty, \infty)$ instead of using the linear model for the response variable of the probability of success on risk factors. The logistic transformation or logit of the probability of success (π) is $\log\{\pi/(1-\pi)\}$, which is written as $\text{logit}(\pi)$ and defined as the log odds of success. It is easily seen that any value of (π) in the range $(0, 1)$ corresponds to the value of $\text{logit}(\pi)$ in $(-\infty, +\infty)$. Usually, binary data results from a nonlinear relationship between $\{\pi(x)\}$ and (x), where a fixed change in (x) has less impact when $\{\pi(x)\}$ is near (0 or 1) than when $\{\pi(x)\}$ is near (0.5). This is illustrated in Figure 1 (see [11]).

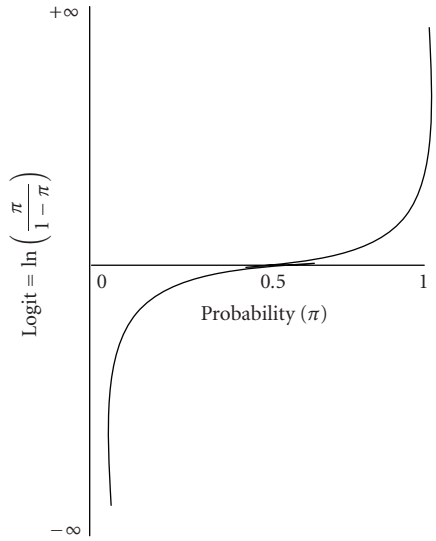

Figure 1: Logit $= \ln(\pi/(1-\pi))$ as a function of the value of probability (π), the logit ranges from $(-\infty)$ to $(+\infty)$ as probability ranges from (0) to (1). The logit = 0 when probability = 0.5.

Thus, the appropriate link is the log odds transformation (the logit). Then if there are n binomial observations of the form $\pi_i = y_i/n_i$ for $i = 1, 2, \ldots, n$, where the expected value of the random variable associated with ith observation, y_i, is $E(Y_i) = n_i\pi_i$. The logistic regression model for association of π_i on the values $x_{1i}, x_{2i}, \ldots, x_{ki}$ of k risk factors $X_1, X_2, \ldots, X_k$ is such that [10]

$$\begin{aligned}\text{Logit}(\pi_i) &= \text{Log}\left\{\frac{\pi_i}{1-\pi_i}\right\} \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki},\end{aligned} \tag{1}$$

and the equation of success probability is

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}. \tag{2}$$

The linear logistic model is a member of a family of generalized linear models (GLM). The next subsection explains this model fitting process.

2.2. Fitting Logistic Regression Models

The mechanics of maximum likelihood (ML) estimation and model fitting for logistic regression model are a special case of GLM fitting, and then fitting the model requires estimation of the unknown parameters (β_j) of the ML function of this model using the Bernoulli ML as in the following [12]:

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}. \tag{3}$$

The problem now is to obtain those values $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ that maximize $\{L(\beta)\}$ or its equivalent $\{\mathrm{Log}\, L(\beta)\}$ where it is as follows:

$$\begin{aligned}
&\mathrm{Log}\, L(\beta)\\
&= \sum_{i=1}^{n}\left\{\mathrm{Log}\binom{n_i}{y_i} + y_i \mathrm{Log}\,\pi_i + (n_i - y_i)\mathrm{Log}(1-\pi_i)\right\}\\
&= \sum_{i=1}^{n}\left\{\mathrm{Log}\binom{n_i}{y_i} + y_i \mathrm{Log}\left(\frac{\pi_i}{1-\pi_i}\right) + (n_i)\mathrm{Log}(1-\pi_i)\right\}\\
&= \sum_{i=1}^{n}\left\{\mathrm{Log}\binom{n_i}{y_i} + y_i\eta_i - (n_i)\mathrm{Log}(1+e^{\eta_i})\right\},
\end{aligned} \tag{4}$$

where $\{\eta_i = \sum_{i=0}^{k}\beta_j x_{ji}\}$ and $(x_{0i} = 1)$ represent all values of (i). The derivative of this log-likelihood function with respect to the $(k+1)$ unknown β-parameters is given by

$$\frac{\partial \mathrm{Log}\, L(\beta)}{\partial(\beta_i)} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} n_i x_{ji} e^{\eta_i}\left(1+e^{\eta_i}\right)^{-1}, \quad j = 0, 1, 2, \ldots, k. \tag{5}$$

Then the likelihood equations are

$$\sum_i y_i x_{ji} - \sum_i n_i \hat{\pi}_i x_{ji} = 0, \quad j = 0, 1, 2, \ldots, k, \tag{6}$$

where $\hat{\pi}_i = e^{\eta_i}(1+e^{\eta_i})^{-1}$ is the ML estimate of $\{\pi_i\}$. There are two methods to solve (6) and obtain the maximum likelihood estimation of $(\hat{\beta})$. The one most often used is known as the Newton-Raphson method. This method begins with determination of the score matrix $\{\mathbf{U}(\beta)\}$ and the information matrix $\{\mathbf{I}(\beta)\}$ as in the following [13]:

$$\begin{aligned}
U_j^{(t)}(\hat{\beta}) &= \left.\frac{\partial L(\beta)}{\partial \beta_j}\right|_{\beta^{(t)}}\\
&= \sum_i (y_i - n_i \pi_i^{(t)}) x_{ij},\\
I_{jk}^{(t)}(\hat{\beta}) &= \left.\frac{\partial^2 L(\beta)}{\partial\beta_j \partial\beta_k}\right|_{\beta^{(t)}}\\
&= -\sum_i x_{ij} x_{ik} n_i \pi_i^{(t)}(1-\pi_i^{(t)}).
\end{aligned} \tag{7}$$

Here $(\pi^{(t)})$ is obtained from $\{\beta^{(t)}\}$ through (2), then we use $\{\mathbf{U}^{(t)}\}$ and $\{\mathbf{I}^{(t)}\}$ with the following formula $\{\beta^{(t+1)} = \beta^{(t)} - (I^{(t)})^{-1}U^{(t)}\}$ to obtain the next value $(\boldsymbol{\beta}^{(t+1)})$ as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \{\mathbf{X}'\mathrm{diag}[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}^{(t)}), \tag{8}$$

where $\{\mu_i^{(t)} = n_i\pi_i^{(t)}\}$, this is to obtain $(\boldsymbol{\pi}^{(t+1)})$, and so on.

2.3. Evaluating the Fitted Model

A simple model that fits adequately has the advantage of model parsimony. If a model has relatively little bias and describes reality well, it tends to provide more accurate estimates of the quantities of interest. Agresti [9] stated that "we are mistaken if we think that we have found the true model, because any model is a simplification of reality." In light of this assertion, what then is the logic of testing the fit of a model when we know that it does not truly hold? The answer lies in the evaluation of the specific properties of this model by using criteria such as deviance, the R^2, the Wald Score test, the Person chi-square, and the Hosmer-Lemshow chi-square tests; for more details, see [9, 10]. Usually the first stage of construction of any model presents a large number of risk factors. Inclusion of all of them may lead to an unattractive model from a statistical viewpoint. Thus, as an important step towards an acceptable model, a decision should be made early about the proper methodology to use to select the appropriate and important risk factors. Because traditional methods of selecting variables have many limitations in their applicability to survival regression models, a new method of variables selection will be developed by using GSA to select the most influential factors in the model. This is the subject of the following section.

3. Sensitivity Analysis to Select the Most Influencing Risk Factors

There are two key problems in variable selection procedure: (i) how to select an appropriate number of risk factors from the set of risk factors, and (ii) how to improve final model performance based on the given data. So answering these questions is the objective of our proposed method by applying GSA to select the influential risk factors in the logistic regression model.

3.1. General Concept of GSA

GSA was defined in [14] as "the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input." Hence one possible way to apportion the importance of the input factors with respect to the model response is to apply GSA. In general the importance of a given risk factor X_i can be measured via the so-called sensitivity index, which is defined as the fractional contribution to the model output variance because of the uncertainty in X_i. For k risk factors, the sensitivity indices can be computed using the following decomposition formula for the total output variance $V(Y)$ of the output Y [15]:

$$V(Y) = \sum_i V(X_i) + \sum_i \sum_{j>i} V(X_i, X_j) + \cdots + V(X_1, \ldots, X_k), \tag{9}$$

where

$$\begin{aligned}
V(X_i) &= V_{x_i}(E_{x_{-i}}(Y \mid X_i)),\\
V(X_i, X_j) &= V_{X_iX_j}(E_{X_{-ij}}(Y \mid X_iX_j)) - V_{X_i}(E_{X_{-i}}(Y \mid X_i))\\
&\quad - V_{X_j}(E_{X_{-j}}(Y \mid X_j)),
\end{aligned} \tag{10}$$

where $V(Y)$ is the unconditional variance of output of the model (incidence of CHD), $V(X_i)$ is the conditional variance of risk factor X_i, and $V(X_i, X_j)$ is the variance of interaction between X_i and X_j, and so on. A first measurement of the fraction of the unconditional output variance $V(Y)$ that is accounted for by the uncertainty in X_i, which is the first-order sensitivity index (S_i) for the factor X_i is given as

$$S_i = \frac{V(X_i)}{V(Y)}. \tag{11}$$

The second terms in (9) are known as the effect of interactions. It is a fact that the number and importance of the interaction terms usually grow (i) with the number of risk factors k, and (ii) with the range of variation of the risk factors [16]. This means that if all of the $V(X_i)$ terms are computed, then most likely $\sum_{i=1}^{k} V(X_i)$ would still be lower than the total $V(Y)$, because the difference $V(Y) - \sum_{i=1}^{k} V(X_i)$ is a measure of the impact of the interactions. Consequently, when $\sum_{i=1}^{k} S_i = 1$, then the model is additive (i.e., without interactions among its input factors), and thus the first order of conditional variances of (10) are all we need to decompose the model output variance. For a nonadditive model, higher-order sensitivity indices account for interaction effects among sets of input factors. However, higher-order sensitivity indices are usually not estimated directly because if the model consists of k risk factors, then the total number of indices (including the S_i's) that should be estimated is as high as $2^k - 1$. For this reason, a more compact sensitivity measurement is used; this measurement is the total effect sensitivity (S_T) index, which concentrates in one single term on all the interactions involving a given factor X_i. For example, for a model of $k = 3$ risk factors, the three total sensitivity indices would be [2]

$$\begin{aligned} S_{T1} &= \frac{V(Y) - V_{X_2X_3}(E_{X_1}(Y \mid X_2X_3))}{V(Y)} \\ &= S_1 + S_{12} + S_{13} + S_{123}, \end{aligned} \tag{12}$$

and analogously

$$\begin{aligned} S_{T2} &= S_2 + S_{12} + S_{23} + S_{123} \\ S_{T3} &= S_3 + S_{13} + S_{23} + S_{123}, \end{aligned} \tag{13}$$

where the conditional variance in (12) expresses the total contribution to the variance of Y because of non-X_i, (i.e., to the $k - 1$ remaining factors), so that $V(Y) - V_{X_{-i}}(E_{X_i}(Y \mid X_{-i}))$ includes all terms (i.e., a first order as well as interactions in (9)) that involve risk factor X_i. For a given risk factor X_i, the coefficient of importance (IC_i) is the difference between S_{Ti} and S_i that reflects an important role of interactions for that risk factor in Y,

$$\mathrm{IC}_i = S_{Ti} - S_i. \tag{14}$$

Explaining the interactions among risk factors helps us to improve our understanding about the model structure. Estimators for both (S_i, S_{Ti}) are provided by a variety of methods such as Sobol, the Fourier amplitude sensitivity test (FAST), and others; for more details, see [17].

3.2. GSA in a Logistic Regression Model

In this study, partitioning the total variance of the objective function $V(Y)$ is the way to estimate S_i and S_{Ti} so as to perform a GSA. How can this model be extended to deal with a binary response variable? Although partitioning of variances is uncomplicated in models with a continuous response variable and a normal error distribution, the extension of this partitioning to models with binary responses is not simple [18]. Consequently, to extend the variance partitioning method to our binary response variable (incident of coronary heart disease (CHD)), suppose that the data is consisting of y_i, the number of people who have CHD. The actual response probability of the incidence of CHD for the ith observation π_i will have a Bernoulli distribution with a mean of p_i, where p_i is the proportion of the patients who have a disease. This response probability is therefore a random variable where $E(\pi_i) = p_i$. The variance of π_i must be equal to zero when p_i is zero or unity, and then a relationship between the unknown probability and our risk factors can be fitted. Typically a logistic regression model represents this relationship between y_i for a sample with n people who have a binomial distribution (i.e., $\{Y_i \sim B(n, \pi_i)\}$, $i = 1, 2, \ldots, n$), and the corresponding response probability of the incidence of the disease is $\pi_i = y_i/n$ for ith observation and the k risk factors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ as in (4) and (5). This model assumes independence between the n observations, and then all the variations conditional on the estimates of the probabilities will be binomial with equal variance:

$$V(Y_i) = n\pi_i(1 - \pi_i). \tag{15}$$

The binomial is not the only possible distribution for fitting proportion data. Other distributions exist that have greater variation (known as overdispersion) or less variation (known as underdispersion) than the binomial distribution conditional on the values of π_i's. The simplest function for the true probability of the ith observation uses a multiplicative scale factor to determine the variance of the response as

$$\mathrm{var}(\pi_i) = r p_i(1 - p_i), \tag{16}$$

where r is a scale factor that is equal to 1. If we have a binomial variation, it will be greater than 1 if there is overdispersion and less than 1 if there is underdispersion, and π_i is an unobservable random variable [19]. The advantages of the multiplicative approach are that it will allow both over- and underdispersions. The random variable Y_i is associated with the observed number of incidences of the disease for the ith unit, y_i. It will have a binomial distribution, and then the mean of Y_i, conditionalon π_i, is

$$E(Y_i \mid \pi_i) = n\pi_i \tag{17}$$

and the conditional variance of Y_i is

$$V(Y_i \mid \pi_i) = rn\pi_i(1 - \pi_i). \tag{18}$$

Since π_i cannot be calculated, then the observed proportion of the disease incidence p_i has to be an estimate of π_i as

$$p_i = \frac{y_i}{n}. \tag{19}$$

According to a standard result from the conditional probability theory, the unconditional expected value of a random variable Y can be obtained from the conditional expectation of Y given X using the equation

$$E(Y) = E\{E(Y \mid X)\} \tag{20}$$

and the unconditional variance of Y is given by [20]

$$V(Y) = E\{V(Y \mid X)\} + V\{E(Y \mid X)\}. \tag{21}$$

Applying these two results on our response variable gives

$$E(Y_i) = E\{E(Y_i \mid \pi_i)\} = E(n\pi_i) = np_i, \tag{22}$$
$$V(Y_i) = E\{V(Y_i \mid \pi_i)\} + V\{E(Y_i \mid \pi_i)\}, \tag{23}$$

now

$$\begin{aligned} E\{V(Y_i \mid \pi_i)\} &= E\{n\pi_i(1-\pi_i)\} \\ &= n\{E(\pi_i) - E(\pi_i^2)\} \\ &= n\{E(\pi_i) - V(\pi_i) - [E(\pi_i)]^2\} \\ &= n\{p_i - rp_i(1-p_i) - p_i^2\} \\ &= np_i(1-p_i)(1-r) \end{aligned} \tag{24}$$

also

$$\operatorname{var}\{E(Y_i \mid \pi_i)\} = \operatorname{var}(n\pi_i) = n^2 r p_i(1-p_i), \tag{25}$$

and so

$$V(Y_i) = nrp_i(1-p_i) \tag{26}$$

in the absence of random variation in the response probability, Y_i would have a binomial distribution, $B(n, p_i)$, and in this case when $r = 1$ as required, then

$$V(Y_i) = np_i(1-p_i). \tag{27}$$

If, on the other hand, r is greater than 1, then a variation in the response probability occurs and the variance of Y_i will exceed $np_i(1-p_i)$, the variance under binomial sampling that leads to overdispersion. But if r is less than 1, then the variation in the response probability and the variance of Y_i will be less than $np_i(1-p_i)$, the variance under binomial sampling that leads to underdispersion. To use GSA to select the important covariates from the available set of covariates and construct an appropriate logistic regression model, it involves three steps.

(1) The first step is identification of the probability distribution $f(x)$ of each covariate in the model. Usually sensitivity analysis starts from probability distribution functions (pdfs) given by the experts. This selection makes the use of the best information available of the statistical properties of the input factors. One of the methods used to obtain the pdfs starts with visualizing the observed data by examining its histogram to see if it is compatible with the shape of any distribution, as illustrated in Figure 2.

A visual approach is not always easy, accurate, or valid, especially if the sample size is small. Thus it would be better to have a more formal procedure for deciding which distribution is "best." A number of significance tests are available for this such as the Kolmogorov-Smirnoff and chi-square tests. For more details, see [21].

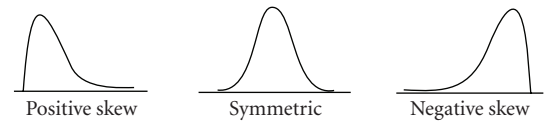

Figure 2: Common shapes of three types of probability distribution.

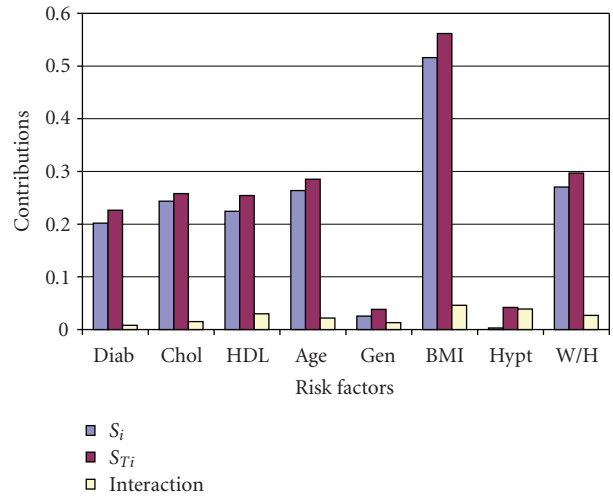

Figure 3: Sensitivity indices: the main effect S_i, the total effect S_{Ti} and the interaction effect IC_i for each risk factor.

(2) In the second step, the logistic regression model as in (1) and the information about the covariates obtained in step one are used to create a Monte Carlo simulation to generate the sample that will be used in the decomposition and to estimate the unconditional variance of response probability and the conditional variation for covariates as in (23) to (26).

(3) These results from step two will be used in performing GSA in the binary logistic regression model using (11), and in the result of decomposing as in (24) and (26), where the main effect indices are

$$S_i = \frac{np_i(1-p_i)(1-r)}{nrp_i(1-p_i)} \tag{28}$$

and the total effect indices are

$$\begin{aligned} S_{Ti} &= \frac{V(Y_j) - V(E(Y_j \mid X_{-i}))}{V(Y_j)} \\ &= \frac{E\{V(Y_j \mid X_{-i})\}}{V(Y_j)}, \end{aligned} \tag{29}$$

where X_{-i} are all X's but X_i, and the coefficients of importance are

$$\mathrm{IC}_i = S_{Ti} - S_i. \tag{30}$$

These results and the two datasets are used to test and compare the performance of the proposed GSA method as

Table 1: Estimated coefficients and standard errors for different variable selection methods.

Methods Factors	MLE	Best subset AIC	Best subset BIC	SCAD	LASSO	SA $S_i(S_{Ti})$
Intercept	5.51 (0.75)	4.81 (0.45)	6.12 (0.57)	6.09 (0.29)	3.70 (0.25)	Constant
X_1	−8.8 (2.97)	−6.49 (1.75)	−12.15 (1.81)	−12.2 (.08)	0 (—)	*0.487 (0.536)*
X_2	2.30 (2.00)	0 (—)	0 (—)	0 (—)	0 (—)	0.014 (0.125)
X_3	−2.77 (3.43)	0 (—)	−6.93 (0.79)	−7.0 (0.21)	0 (—)	*0.143 (0.218)*
X_4	−1.74 (1.41)	0.30 (0.11)	−0.29 (0.11)	0 (—)	−0.28 (0.09)	0.003 (0.034)
X_1^2	−0.75 (.61)	−1.04 (0.54)	0 (—)	0 (—)	−1.71 (0.24)	0.013 (0.057)
X_3^2	−2.7 (2.45)	−4.55 (0.55)	0 (—)	0 (—)	−2.67 (0.22)	0.032 (0.091)
X_1X_2	0.03 (0.34)	0 (—)	0 (—)	0 (—)	0 (—)	0.014 (0.237)
X_1X_3	7.46 (2.34)	5.69 (1.29)	9.83 (1.63)	9.84 (0.14)	0.36 (0.22)	*0.362 (0.502)*
X_1X_4	0.24 (0.32)	0 (—)	0 (—)	0 (—)	0 (—)	0.001 (0.042)
X_2X_3	−2.15 (1.61)	0 (—)	0 (—)	0 (—)	−0.10 (0.10)	0.016 (0.075)
X_2X_4	−0.12 (0.16)	0 (—)	0 (—)	0 (—)	0 (—)	0.003 (0.047)
X_3X_4	1.23 (1.21)	0 (—)	0 (—)	0 (—)	0 (—)	0.019 (0.307)

Table 2: Sensitivity indices and risk factors ranking.

Factors	S_i	S_{Ti}	IC_i	S_i (%)	Ranks
Diab	0.2018	0.22657	0.008	12	**6**
Chol	0.2434	0.258	0.015	14	**4**
HDL	0.2243	0.25424	0.03	13	**5**
Age	0.2636	0.28507	0.022	15	**3**
Gen	0.0256	0.03844	0.013	1	**7**
BMI	0.5161	0.56173	0.046	30	**1**
Hypt	0.003	0.04207	0.039	0	**8**
W/H	0.2706	0.29714	0.027	15	**2**
Sum.	1.7484	1.96326	0.2		

a variable selection method to identify the important risk factors obtained from these datasets with the results obtained from other existing methods of selecting variables.

4. Numerical Comparisons

The purpose of this section is to compare the performance of the proposed method with existing ones. We also use a real data example to illustrate our SA approach as a variable selection method. In the first examples in this section, we used the dataset and the results of the penalized likelihood estimate of best subset (AIC), bust subset (BIC), SCAD, and LASSO that were computed by [7] as a way to compare the performance of the proposed method with these methods.

4.1. The First Example

In this example, Fan and Li [7] applied the proposed penalized likelihood methodology to burn data collected by the General Hospital Burn Center at the University of Southern California. The dataset consists of 981 observations. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise. Risk factors are X_1 = age, X_2 = sex, X_3 = log (burn area + 1), and binary variable X_4 = oxygen (0 normal, 1 abnormal) was considered. Quadratic terms of X_1 and X_3, and all interaction terms were included. The intercept term was added, and the logistic regression model was fitted. The best subset variable selection with the AIC and the BIC was applied to this dataset. The unknown parameter λ was chosen by generalized cross-validation: it is 0.6932 and 0.0015, respectively, for the penalized likelihood estimates with the SCAD and LASSO. The constant a in the SCAD was taken as 3.7. With the selected λ, the penalized likelihood estimator was obtained at the sixth, 28th, and fifth step iterations for the penalized likelihood with the SCAD and LASSO. Table 1 contains the estimated coefficients and standard errors for the transformed data, based on the penalized likelihood estimators, and the calculation of the sensitivity indices obtained by using SimLab software to compare the performance of GSA as a variable selection method with other methods. The first five columns were calculated by [7].

In addition to GSA indices, Table 1 consists of the results of two traditional methods of variable selection (AIC and BIC) and two new methods (LASSO and SCAD). The traditional method, best subset procedure via minimizing the BIC scores, chooses five of 13 risk factors, whereas the SCAD chooses four risk factors. The difference between them is that the best subset keeps X_4. Neither SCAD nor the best subset variable selection (BIC) includes X_1^2 and X_3^2 in the selected subset, but both LASSO and the best subset variable selection (AIC) included them. LASSO chooses the quadratic terms of X_1 and X_3 rather than their linear terms. It also selects an interaction term X_2X_3, which may not be statistically significant. LASSO shrinks noticeably large coefficients. The last column in Table 1 shows that GSA selected the variables X_1, X_3, and X_1X_3, in addition to the intercept, which resembles the SCAD method, and differs from the other methods. According to the results in the last column of Table 1, the risk factors can be ranked according to sensitivity indices S_i and S_{Ti}. Age (X_1) is the first and the most influential risk factor, with a percent of contribution of 0.487, and the second most important risk factor is the interaction

Table 3: The overall fitting criteria for the BEM for a logistic regression model.

Step	$-2\text{Log}\,L$	χ^2_P	Df	Sig.	χ^2_{HL}	df	Sig.	Nag. R^2
6	357.813	7.268	3	0.064	8.465	8	0.389	0.30
7	359.021	6.061	2	0.048	0.055	2	0.973	0.25
8	360.189	4.892	1	0.027	—	—	—	0.20

Table 4: The estimated parameters and their significance for a logistic regression model using BEM.

Steps	Risk factors	$\widehat{\beta}$	Sig. (P)
Step 6	CHOL	0.538	0.061
	SAGE	0.151	0.271
	BMI	−0.325	0.241
	Constant	−1.711	0.000
Step 7	CHOL	0.610	0.029
	BMI	−0.300	0.276
	Constant	−1.758	0.000
Step 8	CHOL	0.605	0.030
	Constant	−1.946	0.000

between X_1 and X_3, with a percentage of contribution of 0.362. The third influential risk factor is the log (area of burn + 1) (X_3) with a percentage of contribution of 0.143 as shown in Table 1. Consequently, we find that the proposed GSA variable selection method resembles SCAD in choosing the same risk factors.

4.2. The Second Example

A new dataset emerges from the original dataset prepared in [22] as a way to compare SA and the traditional method (backward elimination) as variable selection methods. Originally this study was undertaken to determine the prevalence of CHD risk factors among a population-based sample of 403 rural African-Americans in Virginia. Community-based screening evaluations included the determination of exercise and smoking habits, blood pressure, height, weight, total and high-density lipoprotein (HDL) cholesterol, and glycosylated hemoglobin, and other factors. The results of this study were presented as percentages of prevalence for most factors such as diabetes (13.6% of men, 15.6% of women), hypertension (30.9% of men, 43.1% of women), and obesity (38.7% of men, 64.7% of women), without building any models to study the relationship between CHD and its risk factors. For more details, see [8]. A new dataset was generated based on the first one as a way to calculate SA indices to extract the important risk factors for CHD from among these new factors, and then implement the logistic regression model to test the performance of the proposed method as follows.

(1) CHD (Y) 10-year percentage risk is generated according to Framingham Point Scores. This risk is classified as 1 if the percentage of the risk is ≥20% and 0 otherwise [23].

(2) Diabetes (debt, X_1): According to the criteria published by American College of Endocrinology (ACE) & American Association of Clinical Endocrinologists (AACE) [24] the participant has diabetes 1 if the Stabilized Glucose >140 mg/dL or Glycosylated Hemoglobin >7% or both of them more than these limits, and he has no diabetes 0 otherwise.

(3) Total cholesterol (Chol, X_2): if a participant has total cholesterol of >200 mg/dL, he will be given a 1 and a 0 otherwise [25].

(4) High density lipoprotein (HDL, X_3): a participant with HDL of <40 mg/dL will be given a 1 and a 0 otherwise [25].

(5) Age (X_4): standardized values are used $(X - \mu)/\sigma$.

(6) Gender (Gan, X_5): 1 is for a male and 2 for a female.

(7) Body mass index (BMI, X_6): values for this standard are calculated from the following equation: BMI = height/(weight)^2, and the participant gets 1 if BMI is >30 and a 0 otherwise [25].

(8) Blood pressure (hypertension, Hyp, X_7): a participant has Hyp (1) if systolic blood pressure is >140 or if diastolic blood pressure is >90 or if both of them exceed these limits and 0 otherwise [25].

(9) Waist/hip ratio (X_8), in addition to BMI, is a second factor in the determination of obesity.

This dataset was used to perform SA through the use of SimLab software and the partitioning variance methodology discussed in Section 3. An evaluation of the efficiency of the proposed method was performed by fitting all factors into logistic regression models so as to obtain comparisons of factors chosen by the proposed method with those selected by traditional variable selection method (backward elimination). SPSS software was used to get the results that follow from fitting logistic regression models.

4.2.1. The Important Risk Factors

Implementation of the GSA method for this dataset gave the results in Table 2, which shows the ranking of the risk factors in order of importance and the contribution of each one to explaining the total variance of the CHD response variable.

According to the first order of sensitivity indices S_i, the BMI is the first and the most influential factor, and the waist-hip ratio ranks second. Both are components of the obesity factor. Age is the third influential factor and so on through the other factors as listed in Table 2. The total sensitivity index for a given risk factor provides a measure of the overall contribution of that risk factor to the output variance, taking

into account all possible interactions with the other risk factors. The difference between the total sensitivity index and the first-order sensitivity index for a given risk factor is a measure of the contribution to the output variance of the interactions involving that factor; see (12) and (13). The second column in Table 2 shows the values of S_{Ti}, which gives the same rank as S_i for the risk factors. These indices point to the simple interaction between these risk factors as illustrated in the third column in the same table. Figure 3 shows the compression between the first order S_i, the total S_{Ti} sensitivity indices, and the interactions between risk factors.

4.2.2. Implementing the Logistic Regression Model

Does the proposed method yield a reliable model? To investigate the reliability of the proposed method, we compared the results of the fitted models. Basically, when the full logistic regression model is fitted, the results are

$$\begin{aligned}\text{Logit CHD} &= 0.365 - 0.266\,\text{Diab} + 0.557\,\text{Chol} - 0.246\,\text{HDL} \\ &\quad + 0.161\,\text{Age} - 0.147\,\text{Gend} - 0.295\,\text{BMI} \\ &\quad + 0.024\,\text{Hypt} - 1.874\,\text{W/H ratio} \\ \text{Sig}\,(P) &\quad (0.862)\quad (0.480)\quad (0.054)\quad (0.419) \\ &\quad (0.304)\quad (0.624)\quad (0.317) \\ &\quad (0.935)\quad (0.389),\end{aligned} \tag{31}$$

$$\begin{aligned}-2\log L_0 &= 365.081, \quad -2\log L_f = 355.687, \\ \text{Nag. } R^2 &= 0.39, \quad \chi_P^2 = 9.394, \quad \text{Sig. } (P) = 0.310, \\ \chi_{HL}^2 &= 12.509, \quad \text{Sig. } (P) = 0.130,\end{aligned} \tag{32}$$

These results showed the significance of the overall fit of the model according to the values of χ_P^2 and χ_{HL}^2 in spite of the low value of Nag. R^2; also showed that the individual effect for all risk factors is not significant, which means that H_0 cannot be rejected from the following null hypothesis:

$$H_0 : \widehat{\beta} = 0 \quad \text{versus} \quad H_1 : \widehat{\beta} \neq 0. \tag{33}$$

Second, application of the logistic regression model by using those risk factors that appear in Table 2 as highly ranked by the proposed method also shows that this method ranks each risk factor according to its contribution to the incidence of the CHD response variable. The question also becomes how many variables must be selected in order to apply the logistic regression model. The possibility exists that the selection procedure may tend to underfit or overfit the model by selecting too few or too many variables. In the face of such a possibility, our objective becomes to find the model that uses the least number of variables while simultaneously explaining a reasonable percentage of variance in the dependent variable relative to the percentage explained by all the variables in the full model. Thus two models may be fitted from Table 2 to compare the results. The first logistic regression model consisted of the obesity factors (BMI, and W/H ratio), age, and total cholesterol factors that explained 74% of the total variance of the CHD response variable according to the individual effect (S_i) as in Table 2. The results of fitting this model in this manner and applying SPSS software were

$$\begin{aligned}\text{Logit CHD} &= -0.866 + 0.537\,\text{Chol} + 0.170\,\text{Age} \\ &\quad - 0.352\,\text{BMI} - 0.939\,\text{W/H ratio} \\ \text{Sig}\,(P) &\quad (0.026)\quad (0.024)\quad (0.023) \\ &\quad (0.021)\quad (0.036),\end{aligned} \tag{34}$$

$$\begin{aligned}-2\log L_0 &= 365.081, \quad -2\log L_R = 357.584, \\ \text{Nag. } R^2 &= 0.71, \quad \chi_P^2 = 7.497, \quad \text{Sig. } (P) = 0.112, \\ \chi_{HL}^2 &= 16.791, \quad \text{Sig. } (P) = 0.320.\end{aligned} \tag{35}$$

The results in (34) showed that using these criteria for the overall fit for this model demonstrated their significance collectively and individually as risk factors that influence the incidence of CHD and raise the value of R^2 to 71% in comparison with the full model in (31) The second logistic regression model is fitted by adding another risk factor, HDL, to increase the percentage of explanation to 87%. The results of fitting this model as in (36) are

$$\begin{aligned}\text{Logit CHD} &= -0.331 + 0.552\,\text{Chol} - 0.316\,\text{HDL} + 0.175\,\text{Age} \\ &\quad - 0.306\,\text{BMI} - 1.351\,\text{W/H} \\ \text{Sig. } (P) &\quad (0.085)\quad (0.056)\quad (0.28)\quad (0.022) \\ &\quad (0.028)\quad (0.05),\end{aligned} \tag{36}$$

$$\begin{aligned}-2\log L_{1\text{st}} &= 357.584, \quad -2\log L_{2\text{nd}} = 356.434, \\ \text{Nag. } R^2 &= 0.698, \quad \chi_P^2 = 8.648, \quad \text{Sig. } (P) = 0.124, \\ \chi_{HL}^2 &= 4.850, \quad \text{Sig. } (P) = 0.773.\end{aligned} \tag{37}$$

These results showed that adding the HDL risk factor does not improve the results of the first logistic regression model, but the parameter of this risk factor is not significant when we test the following hypothesis:

$$H_0 : \beta_{\text{HDL}} = 0 \quad \text{versus} \quad H_1 : \beta_{\text{HDL}} \neq 0. \tag{38}$$

Note that the difference between the deviances of the two models is minor. Furthermore, the value of R^2 does not improve. Thus, according to the principle of parsimony, the first model should be considered the best model and the risk factors used to construct this model are those that are the most influential in causing CHD. Moreover, showing the different results obtained from these two models demonstrates the differences between fitting the full model with all risk factors and fitting it with only selected risk factors.

The efficiency of the proposed method of variable selection (GSA) can be measured by comparing its results as in (34) with the results gained from fitting the logistic regression model by using the backward elimination method (BEM). These results are shown in Tables 3 and 4.

Table 3 shows the overall fitting criteria required for the last three steps of a logistic regression model fitted by the use of the BEM.

Also Table 4 shows the last three steps of iteration to choose the important risk factors. These results represent the sequential elimination of the factors, which requires eight steps to rank these risk factors according to their importance; however, the proposed method does not need these iterations.

5. Conclusions

The results in Tables 1 to 4 and (31) to (36) for the two examples confirm that the proposed method is capable of distinguishing between important and unimportant risk factors. The proposed method ranked the risk factors according to their decreasing importance as shown in Tables 1 and 2. In the example in which we compared the proposed method with those methods that are typically used, we found that its performance very much resembled the SCAD method in which the same risk factors are selected. From the first example, we found that the important risk factors are age, the area of the burns, and the interaction between them. In the second example we found that the obesity factors (BMI and W/H) are the most influential risk factor on the incidence of CHD, the second risk factor is age, and the third risk factor is the total cholesterol. These play the major roles, representing approximately 74% of the incidence of CHD. Thus, they are considered the most important risk factors according to their individual percentages of contribution in the incidence of CHD as shown in Table 1. Compression between the results of the fitting of the full logistic regression model as in (31) and the chosen models as in (34) and (36) confirm the efficiency of the proposed method in its selection of the most important risk factors. Equation (34) represents the best model, according to the model evaluation criteria, because it consists of the most influential risk factors. Therefore, a medical care plan and medical interventions should comply with this ordering of these factors. Also, to further confirm these results, one of the traditional variable selection methods was used (backward elimination method), which yields different results after eight steps, but the proposed method orders the risk factors without iteration and without the need to fit multiple regression models. Finally, these results together confirm and emphasize the importance of GSA as a variable selection method.

Acknowledgment

This work was supported by USM fellowship.

References

[1] A. Saltelli, K. Chan, and E. M. Scott, *Sensitivity Analysis*, John Wiley & Sons, Chichester, UK, 2000.

[2] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, "Sensitivity analysis for chemical models," *Chemical Reviews*, vol. 105, no. 7, pp. 2811–2827, 2005.

[3] A. Khalili and J. Chen, "Variable selection in finite mixture of regression models," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 1025–1038, 2007.

[4] A. J. Miller, *Subset Selection in Regression*, Chapman & Hall/CRC, London, UK, 2nd edition, 2002.

[5] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[6] J. Fan and R. Li, "Variable selection for Cox's proportional hazards model and frailty model," *Annals of Statistics*, vol. 30, no. 1, pp. 74–99, 2002.

[7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[8] H. H. Zhang and W. Lu, "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, 2007.

[9] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2002.

[10] D. Collett, *Modeling Binary Data*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2nd edition, 2003.

[11] J. Cohen, P. Cohen, S. G. West, and L. S. Alken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 3rd edition, 2003.

[12] D. R. Cox and E. J. Snell, *Analysis of Binary Data*, Chapman & Hall/CRC, New York, NY, USA, 2nd edition, 1989.

[13] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer, New York, NY, USA, 2000.

[14] A. Saltelli, "Global sensitivity analysis: an introduction," in *Sensitivity Analysis of Model Output*, K. M. Hanson and F. M. Hemez, Eds., pp. 27–43, Los Alamos National Laboratory, Los Alamos, NM, USA, 2005.

[15] A. Saltelli, S. Tarantola, and K. P.-S. Chan, "A quantitative model-independent method for global sensitivity analysis of model output," *Technometrics*, vol. 41, no. 1, pp. 39–56, 1999.

[16] A. Saltelli, S. Tarantola, and F. Campolongo, "Sensitivity analysis as an ingredient of modeling," *Statistical Science*, vol. 15, no. 4, pp. 377–395, 2000.

[17] K. Chan, S. Tarantola, A. Saltelli, and I. M. Sobol', "Variance based methods," in *Sensitivity Analysis*, A. Saltelli, K. Chan, and M. Scott, Eds., pp. 167–197, John Wiley & Sons, New York, NY, USA, 2000.

[18] J. Neter, H. K. Michael, J. N. Christopher, and W. William, *Applied Linear Statistical Models*, McGraw-Hill, New York, NY, USA, 1996.

[19] J. S. Long, *Regression Models for Categorical and Limited Dependent Variables*, Sage, Thousand Oaks, Calif, USA, 1997.

[20] M. Saisana, A. Saltelli, and S. Tarantola, "Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators," *Journal of the Royal Statistical Society. Series A*, vol. 168, no. 2, pp. 307–323, 2005.

[21] A. Heiat, "Using an Excel extension for selecting the probability distribution of empirical data," *Spreadsheets in Education*, vol. 2, no. 1, pp. 95–100, 2005.

[22] J. B. Schorling, J. Roach, M. Siegel, et al., "A trial of church-based smoking cessation interventions for rural African Americans," *Preventive Medicine*, vol. 26, no. 1, pp. 92–101, 1997.

[23] J. I. Cleeman, S. M. Grundy, D. Becker, et al., "Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III)," *The Journal of the American Medical Association*, vol. 285, no. 19, pp. 2486–2497, 2001.

[24] J. T. DiPiro, R. L. Talbert, G. C. Yee, G. R. Matzke, B. G. Wells, and L. M. Posey, *Pharmacotherapy: A Pathophysiologic Approach*, McGraw-Hill, New York, NY, USA, 6th edition, 2005.

[25] M. A. Koda-Kimble, L. Y. Young, W. A. Kradian, B. J. Guglielmo, B. K. Allderege, and R. L. Corelli, *Applied Therapeutics, The Clinical Use of Drugs*, Lippincott Williams & Wilkins, Baltimore, Md, USA, 8th edition, 2005.