Hindawi Publishing Corporation

International Journal of Quality, Statistics, and Reliability

International Journal of Quality, Statistics, and Reliab
Volume 2008 (2008), Article ID 471607, 10 pages
doi:10.1155/2008/471607

Research Article

Sensitivity Analysis to Select the Most I Risk Factors in a Logistic Regression Mo

Jassim N. Hussain

School of Mathematical Sciences, University Sains Malaysia, 11800

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

Academic Editor: Myong K. (MK) Jeong

Abstract

The traditional variable selection methods for survival data dep process assumes tuning parameters that are problematic and tim and have a large number of risk factors. In this paper, we prop analysis (GSA) to select the most influential risk factors. This co model by excluding the irrelevant risk factors, thus eliminating the Data from medical trials are suggested as a way to test the efficie simplify the model. This leads to construction of an appropriate n according to their importance.

1. Introduction

Sensitivity analysis (SA) plays a central role in a variety of sta discrimination, calibration, comparison, and model selection [1]. input factors (if any) accounts for most of the output variance (a percentage can be fixed to any value within their range [2]. In important variables to simplification of the model; the original mo best arrive at such a determination. Although SA has been wic important input variables from a complex model so as to arrive at

it has limited use for selection of risk factors despite the prese
regression models. The limited use of these methods to select
illustrates the desirability of development of a new method of SA-
traditional methods and also simplify survival regression models by

A considerable number of methods of variable selection have b
developments are squarely in the context of normal regression m
linear regression models [3]. A comprehensive review of many
Methods such as forward, backward, and stepwise selection and s
and Bayesian information criterion (BIC)) are available; however n
in either a logistic regression model or in other survival regress
standard errors and *P*-values. They also can delete variables whose
regard all the risk factors of a situation as equal, and they se
sequentially; furthermore, most of these methods focus on th
(interactions of variables).

New methods of variable selection, such as *least absolute shrinka*
smoothly clipped absolute deviation (SCAD) method in [6], are
survival regression models. These methods use the penalized li
approaches. These two approaches differ from traditional methods
the model by estimating their effects as 0. A nice feature of th
variable selection simultaneously, but, nevertheless, these metho
problems that are dealt with in more detail in [7, 8].

This study aims to use SA to extend and develop an effective, effi
which the best subsets are identified according to specified criteri
regression models in the field of survival regression models. The
Section 2 gives the background of building a logistic regressio
method. The results of implementing this method and logistic re
Section 5 consists of the discussion and conclusions.

2. Background of Constructing a Logistic Regressio

Often the response variable in clinical data is not a numerical valu
not diseased). When the latter occurs, a binary logistic regressio
relationship between the disease' s measurements and its risk
response variable (the disease measurement) is a dichotomy and t
logistic regression model neither assumes the linearity in the rela
variable, nor does it require normally distributed variables. It also
has less stringent requirements than linear regression models.
independent and that the independent risk factors are linearly r
However, models involving the association between risk factors a
disciplines such as medicine, engineering, and the natural science
factors and binary response variable? The answer to this question i

2.1. Constructing a Logistic Regression Model

The first step in modeling binomial data is a transformation of the
of using the linear model for the response variable of the pr
transformation or logit of the probability of success (π) is $\log\{\pi /$
the log odds of success. It is easily seen that any value of (π) in th
$(-\infty, +\infty)$. Usually, binary data results from a nonlinear relationship

has less impact when $\{\pi(x)\}$ is near (0 or 1) than when $\{\pi(x)\}$ is n

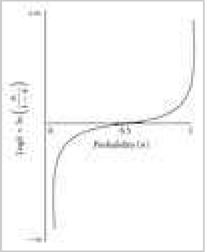

Figure 1: $\text{Logit} = \ln(\pi/(1-\pi))$ as a function o $(-\infty)$ to $(+\infty)$ as probability ranges from (0) to (:

Thus, the appropriate link is the log odds transformation (the logi form $\pi_i = y_i/n_i$ for $i = 1, 2, \ldots, n$, where the expected value of the r is $E(Y_i) = n_i\pi_i$. The logistic regression model for association of $X_1, X_2, \ldots, X_k$ is such that [10]

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

and the equation of success probability is

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdot}$$

The linear logistic model is a member of a family of generalized this model fitting process.

2.2. Fitting Logistic Regression Models

The mechanics of maximum likelihood (ML) estimation and mode case of GLM fitting, and then fitting the model requires estimation of this model using the Bernoulli ML as in the following [12]:

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

The problem now is to obtain those values $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)$ that max as follows:

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

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

$$\frac{\partial \mathrm{Log}\,L(\beta)}{\partial(\beta_i)} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} n_i x$$

$$j = 0, 1, 2, \ldots ,$$

Then the likelihood equations are

$$\sum_i y_i x_{ji} - \sum_i n_i \tilde{\pi}_i x_{ji} = 0, \quad j =$$

where $\tilde{\pi}_i = e^{\eta_i}(1+e^{\eta_i})^{-1}$ is the ML estimate of $\{\pi_i\}$. There are tv likelihood estimation of $(\tilde{\beta})$. The one most often used is known as with determination of the score matrix $\{\mathbf{U}(\beta)\}$ and the information

$$U_j^{(t)}(\tilde{\beta}) = \left.\frac{\partial L(\beta)}{\partial \beta_j}\right|_{\beta^{(t)}}$$

$$= \sum_i (y_i - n_i \pi_i^{(t)}) x$$

$$I_{jk}^{(t)}(\tilde{\beta}) = \left.\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}\right|_{\beta^{(t)}}$$

$$= -\sum_i x_{ij} x_{ik} n_i \pi_i^{(t}$$

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

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \{\mathbf{X}'\mathrm{diag}[n_i \pi_i^{(t)}(1-\pi$$

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 describes reality well, it tends to provide more accurate estimates "we are mistaken if we think that we have found the true model In light of this assertion, what then is the logic of testing the fit of

The answer lies in the evaluation of the specific properties of this the Wald Score test, the Person chi-square, and the Hosmer-Lems Usually the first stage of construction of any model presents a la may lead to an unattractive model from a statistical viewpoint. model, a decision should be made early about the proper methodc risk factors. Because traditional methods of selecting variables ha regression models, a new method of variables selection will be de factors in the model. This is the subject of the following section.

3. Sensitivity Analysis to Select the Most Influencin

There are two key problems in variable selection procedure: (i) h from the set of risk factors, and (ii) how to improve final model p these questions is the objective of our proposed method by apply logistic regression model.

3.1. General Concept of GSA

GSA was defined in [14] as “the study of how the uncertainty in be apportioned to different sources of uncertainty in the model importance of the input factors with respect to the model respons given risk factor X_i can be measured via the so-called sensitivity i to the model output variance because of the uncertainty in X_i. computed using the following decomposition formula for the total c

$$V(Y)=\sum_i V(X_i)+\sum_i \sum_{j>i} V(X_i,X_j)+$$

where

$$V(X_i)=V_{X_i}(E_{X_{-i}}(Y\,|$$

$$\begin{aligned} V(X_i,X_j) &= V_{X_iX_j}(E_{X_{-ij}}(Y\,|\,X_iX_j)) \\ &\quad -V_{X_j}(E_{X_{-j}}(Y\,|\,X_j)), \end{aligned}$$

where $V(Y)$ is the unconditional variance of output of the model (in risk factor X_i, and $V(X_i,X_j)$ is the variance of interaction between fraction of the unconditional output variance $V(Y)$ that is account order sensitivity index (S_i) for the factor X_i is given as

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

The second terms in (9) are known as the effect of interactions. interaction terms usually grow (i) with the number of risk factor factors [16]. This means that if all of the $V(X_i)$ terms are compute than the total $V(Y)$, because the difference $V(Y)-\sum_{i=1}^{k} V(X_i)$ is Consequently, when $\sum_{i=1}^{k} S_i=1$, then the model is additive (i.e., thus the first order of conditional variances of (10) are all we nee nonadditive model, higher-order sensitivity indices account for However, higher-order sensitivity indices are usually not estimat factors, then the total number of indices (including the S_i's) that reason, a more compact sensitivity measurement is used; this me which concentrates in one single term on all the interactions invol $k=3$ risk factors, the three total sensitivity indices would be [2]

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

and analogously

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

where the conditional variance in (12) expresses the total contribu the $k-1$ remaining factors), so that $V(Y)-V_{X_{-i}}(E_{X_i}(Y\,|\,X_{-i}))$ include

in (9)) that involve risk factor X_i. For a given risk factor X_i, thecoe
S_{Ti} and S_i that reflects an important role of interactions for that ris

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

Explaining the interactions among risk factors helps us to impro
Estimators for both (S_i, S_{Ti}) are provided by a variety of methods s
(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 funct
perform a GSA. How can this model be extended to deal with a
variances is uncomplicated in models with a continuous respor
extension of this partitioning to models with binary responses
variance partitioning method to our binary response variable (incid
the data is consisting of y_i, the number of people who have CHD.
CHD for the ith observation π_i will have a Bernoulli distribution w
patients who have a disease. This response probability is therefor
of π_i must be equal to zero when p_i is zero or unity, and then a re
risk factors can be fitted. Typically a logistic regression model re
with *n* people who have a binomial distribution (i.e., $\{Y_i \sim B(n,\pi$
probability of the incidence of the disease is $\pi_i = y_i / n$ for ith obser
and (5). This model assumes independence between the *n* observa
estimates of the probabilities will be binomial with equal variance:

$$V(Y_i) = n\pi_i(1 - \pi$$

The binomial is not the only possible distribution for fitting proport
variation (known as overdispersion) or less variation (known a
conditional on the values of π_i's. The simplest function for th
multiplicative scale factor to determine the variance of the respons

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

where *r* is a scale factor that is equal to 1. If we have a binor
overdispersion and less than 1 if there is underdispersion, and
advantages of the multiplicative approach are that it will allow bot
Y_i is associated with the observed number of incidences of the c
distribution, and then the mean of Y_i, conditionalon π_i, is

$$E(Y_i \mid \pi_i) = n\pi_i$$

and the conditional variance of Y_i is

$$V(Y_i \mid \pi_i) = rn\pi_i(1 -$$

Since π_i cannot be calculated, then the observed proportion of the

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

According to a standard result from the conditional probability the
variable *Y* can be obtained from the conditional expectation of *Y* gi

$$E(Y) = E\{E(Y \mid X$$

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

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

Applying these two results on our response variable gives

$$E(Y_i) \;= E\{E(Y_i \mid n_i)\} = E($$

$$V(Y_i) \;= E\{V(Y_i \mid n_i)\} + V\{$$

now

$$\begin{aligned} E\{V(Y_i \mid n_i)\} &= E\{nn_i(1 - n_i \\ &= n\{E(n_i) - E(\\ &= n\{E(n_i) - V(\\ &= n\{p_i - rp_i(1 \\ &= np_i(1 - p_i)(1 \end{aligned}$$

also

$$\operatorname{var}\{E(Y_i \mid n_i)\} = \operatorname{var}(nn_i) =$$

and so

$$V(Y_i) = nrp_i(1 -$$

in the absence of random variation in the response probability, Y_i this case when $r = 1$ as required, then

$$V(Y_i) = np_i(1 - p_i$$

If, on the other hand, r is greater than 1, then a variation in the re exceed $np_i(1 - p_i)$, the variance under binomial sampling that lead variation in the response probability and the variance of Y_i will be sampling that leads to underdispersion. To use GSA to select th covariates and construct an appropriate logistic regression model,

(1) The first step is identification of the probability distribution $f(x$ analysis starts from probability distribution functions (pdfs) given best information available of the statistical properties of the input starts with visualizing the observed data by examining its histogr distribution, as illustrated in Figure 2.

Figure 2: Common shapes of three types of pr

A visual approach is not always easy, accurate, or valid, especially to have a more formal procedure for deciding which distributio available for this such as the Kolmogorov-Smirnoff and chi-square

(2) In the second step, the logistic regression model as in (1) an step one are used to create a Monte Carlo simulation to generate and to estimate the unconditional variance of response probabilit (23) to (26).

(3) These results from step two will be used in performing GSA in

in the result of decomposing as in (24) and (26), where the main e

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

and the total effect indices are

$$S_{Ti} = \frac{V(Y_j)-V(E(Y_j}{V(Y_j)} = \frac{E\{V(Y_j|X_{-i})\}}{V(Y_j)}$$

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

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

These results and the two datasets are used to test and compare variable selection method to identify the important risk factors obt from other existing methods of selecting variables.

4. Numerical Comparisons

The purpose of this section is to compare the performance of the real data example to illustrate our SA approach as a variable sele we used the dataset and the results of the penalized likelihood SCAD, and LASSO that were computed by [7] as a way to comp these methods.

4.1. The First Example

In this example, Fan and Li [7] applied the proposed penalized lik General Hospital Burn Center at the University of Southern Califor binary response variable Y is 1 for those victims who survived the age, X_2 = sex, X_3 = log (burn area + 1), and abnormal) was considered. Quadratic terms of X_1 and X_3, and all i was added, and the logistic regression model was fitted. The best was applied to this dataset. The unknown parameter λ was chose 0.0015, respectively, for the penalized likelihood estimates with t was taken as 3.7. With the selected λ, the penalized likelihood es step iterations for the penalized likelihood with the SCAD and LAS standard errors for the transformed data, based on the penalize sensitivity indices obtained by using SimLab software to compar method with other methods. The first five columns were calculated

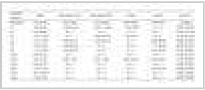

Table 1: Estimated coefficients and standard e

In addition to GSA indices, Table 1 consists of the results of two BIC) and two new methods (LASSO and SCAD). The traditional me scores, chooses five of 13 risk factors, whereas the SCAD choose that the best subset keeps X_4. Neither SCAD nor the best subset selected subset, but both LASSO and the best subset variable s quadratic terms of X_1 and X_3 rather than their linear terms. It also statistically significant. LASSO shrinks noticeably large coefficier selected the variables X_1, X_3, and X_1X_3, in addition to the interce from the other methods. According to the results in the last cc according to sensitivity indices S_i and S_{Ti}. Age (X_1) is the first an contribution of 0.487, and the second most important risk fact percentage of contribution of 0.362. The third influential risk f percentage of contribution of 0.143 as shown in Table 1. Conse selection method resembles SCAD in choosing the same risk factor

4.2. The Second Example

A new dataset emerges from the original dataset prepared in [22] (backward elimination) as variable selection methods. Original prevalence of CHD risk factors among a population-based sar Community-based screening evaluations included the determinatic height, weight, total and high-density lipoprotein (HDL) cholesterc The results of this study were presented as percentages of preva men, 15.6% of women), hypertension (30.9% of men, 43.1% o women), without building any models to study the relationship b see [8]. A new dataset was generated based on the first one as a risk factors for CHD from among these new factors, and then in performance of the proposed method as follows.

(1) CHD (Y) 10-year percentage risk is generated according as 1 if the percentage of the risk is ≥20% and 0 otherwise [23]

(2) Diabetes (debt, X_1): According to the criteria publishe American Association of Clinical Endocrinologists (AACE) [24 Glucose >140 mg/dL or Glycosylated Hemoglobin >7% or bo diabetes 0 otherwise.

(3) Total cholesterol (Chol, X_2): if a participant has total cho 0 otherwise [25].

(4) High density lipoprotein (HDL, X_3): a participant with 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 a height/(weight)^2, and the participant gets 1 if BMI is >30 and a

(8) Blood pressure (hypertension, Hyp, X_7): a participant h diastolic blood pressure is >90 or if both of them exceed these

(9) Waist/hip ratio (X_8), in addition to BMI, is a second facto

This dataset was used to perform SA through the use of SimLab s discussed in Section 3. An evaluation of the efficiency of the propo logistic regression models so as to obtain comparisons of factors c by traditional variable selection method (backward elimination). SI from fitting logistic regression models.

4.2.1. The Important Risk Factors

Implementation of the GSA method for this dataset gave the resu factors in order of importance and the contribution of each one tc variable.

Table 2: Sensitivity indices and risk factors ra

According to the first order of sensitivity indices S_i, the BMI is the hip ratio ranks second. Both are components of the obesity factor. the other factors as listed in Table 2. The total sensitivity index overall contribution of that risk factor to the output variance, tal other risk factors. The difference between the total sensitivity inc risk factor is a measure of the contribution to the output variance and (13). The second column in Table 2 shows the values of S_{Ti}, v These indices point to the simple interaction between these risk fa table. Figure 3 shows the compression between the first order S_i, between risk factors.

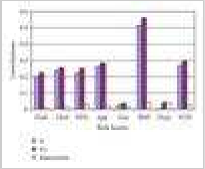

Figure 3: Sensitivity indices: the main effect IC_i for each risk factor.

4.2.2. Implementing the Logistic Regression Model

Does the proposed method yield a reliable model? To investigate t the results of the fitted models. Basically, when the full logistic reg

$$\begin{aligned} \text{Logit CHD } &= 0.365 - 0.266\,\text{Diab} + 0 \\ &+0.161\,\text{Age} - 0.147\,\text{Ge} \\ &+0.024\,\text{Hypt} - 1.874\,\text{W} \\ \text{Sig}\,(P) \quad & (0.862) \qquad (0.480) \\ & \quad (0.304) \qquad (0.624) \\ & \quad (0.935) \qquad\quad (0.38 \end{aligned}$$

$$-2\log L_0 = 365.081, \qquad -2\log$$

$$\text{Nag.}R^2 = 0.39, \qquad \chi^2_P = 9.394,$$

$$\chi^2_{HL} = 12.509, \qquad \text{Sig.}\,(P$$

These results showed the significance of the overall fit of the mod the low value of $\text{Nag.}R^2$; also showed that the individual effect for H_0 cannot be rejected from the following null hypothesis:

$$H_0 : \vec{\beta} = 0 \quad \text{versus} \quad H$$

Second, application of the logistic regression model by using those by the proposed method also shows that this method ranks ea incidence of the CHD response variable. The question also become apply the logistic regression model. The possibility exists that the the model by selecting too few or too many variables. In the face the model that uses the least number of variables while simul variance in the dependent variable relative to the percentage expla models may be fitted from Table 2 to compare the results. The firs factors (BMI, and W/H ratio), age, and total cholesterol factors th response variable according to the individual effect (S_i) as in Table and applying SPSS software were

$$\begin{aligned} \text{Logit CHD} &= -0.866 + 0.537\,\text{C} \\ &\quad -0.352\,\text{BMI} - 0.9 \\ \text{Sig}\,(P) &\qquad (0.026) \qquad (0.0 \\ &\qquad\quad (0.021) \end{aligned}$$

$$\begin{aligned} -2\log L_0 &= 365.081, \qquad -2\log \\ \text{Nag.}\,R^2 = 0.71, &\qquad \chi^2_P = 7.497, \\ \chi^2_{HL} &= 16.791, \qquad \text{Sig.}\,(P \end{aligned}$$

The results in (34) showed that using these criteria for the overa collectively and individually as risk factors that influence the incic comparison with the full model in (31) The second logistic regres HDL, to increase the percentage of explanation to 87%. The results

$$\begin{aligned} \text{Logit CHD} &= -0.331 + 0.552\,\text{Chol} - 0 \\ &\quad -0.306\,\text{BMI} - 1.351\,\text{W}, \\ \text{Sig.}\,(P) &\qquad (0.085) \quad (0.056) \\ &\qquad (0.028) \qquad (0.05), \end{aligned}$$

$$\begin{aligned} -2\log L_{1\text{st}} &= 357.584, \qquad -2\log \\ \text{Nag.}\,R^2 = 0.698, &\qquad \chi^2_P = 8.648, \\ \chi^2_{HL} &= 4.850, \qquad \text{Sig.}\,(P) \end{aligned}$$

These results showed that adding the HDL risk factor does not model, but the parameter of this risk factor is not significant when

$$H_0 : \beta_{\text{HDL}} = 0 \quad \text{versus} \quad H$$

Note that the difference between the deviances of the two mode improve. Thus, according to the principle of parsimony, the first n risk factors used to construct this model are those that are the n the different results obtained from these two models demonstrates all risk factors and fitting it with only selected risk factors.

The efficiency of the proposed method of variable selection (GSA (34) with the results gained from fitting the logistic regression (BEM). These results are shown in Tables 3 and 4.

Table 3: The overall fitting criteria for the BEM

Table 4: The estimated parameters and their
BEM.

Table 3 shows the overall fitting criteria required for the last thre
use of the BEM.

Also Table 4 shows the last three steps of iteration to choose the
sequential elimination of the factors, which requires eight ste
importance; however, the proposed method does not need these it

5. Conclusions

The results in Tables 1 to 4 and (31) to (36) for the two exampl
distinguishing between important and unimportant risk factors
according to their decreasing importance as shown in Tables 1 a
proposed method with those methods that are typically used, we f
SCAD method in which the same risk factors are selected. From t
factors are age, the area of the burns, and the interaction betwe
obesity factors (BMI and W/H) are the most influential risk factor
age, and the third risk factor is the total cholesterol. These play t
the incidence of CHD. Thus, they are considered the most im
percentages of contribution in the incidence of CHD as shown in
fitting of the full logistic regression model as in (31) and the chose
of the proposed method in its selection of the most important risk
according to the model evaluation criteria, because it consists of th
care plan and medical interventions should comply with this orde
results, one of the traditional variable selection methods was u
different results after eight steps, but the proposed method order
need to fit multiple regression models. Finally, these results toget
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*, Jo

2. A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, "Se
Reviews, vol. 105, no. 7, pp. 2811–2827, 2005.

3. A. Khalili and J. Chen, "Variable selection in finite mixture c
Statistical Association, vol. 102, no. 479, pp. 1025–1038, 2

4. A. J. Miller, *Subset Selection in Regression*, Chapman & Hall,

5. R. Tibshirani, "The lasso method for variable selection in th
4, pp. 385–395, 1997.

6. J. Fan and R. Li, “Variable selection for Cox's proportional h *Statistics*, vol. 30, no. 1, pp. 74–99, 2002.

7. J. Fan and R. Li, “Variable selection via nonconcave penaliz *the American Statistical Association*, vol. 96, no. 456, pp. 13

8. H. H. Zhang and W. Lu, “Adaptive Lasso for Cox's proportio pp. 691–703, 2007.

9. A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, Ho

10. D. Collett, *Modeling Binary Data*, Chapman & Hall/CRC, Boca

11. J. Cohen, P. Cohen, S. G. West, and L. S. Alken, *Applied Mu* *Behavioral Sciences*, Lawrence Erlbaum Associates, Mahwah

12. D. R. Cox and E. J. Snell, *Analysis of Binary Data*, Chapman

13. T. M. Therneau and P. M. Grambsch, *Modeling Survival Data* NY, USA, 2000.

14. A. Saltelli, “Global sensitivity analysis: an introduction,” in and F. M. Hemez, Eds., pp. 27–43, Los Alamos National Lab

15. A. Saltelli, S. Tarantola, and K. P.-S. Chan, “A quantitative analysis of model output,” *Technometrics*, vol. 41, no. 1, p

16. A. Saltelli, S. Tarantola, and F. Campolongo, “Sensitivity an *Science*, vol. 15, no. 4, pp. 377–395, 2000.

17. K. Chan, S. Tarantola, A. Saltelli, and I. M. Sobol', “Varianc Saltelli, K. Chan, and M. Scott, Eds., pp. 167–197, John Wil

18. J. Neter, H. K. Michael, J. N. Christopher, and W. William, *A* York, NY, USA, 1996.

19. J. S. Long, *Regression Models for Categorical and Limited De* USA, 1997.

20. M. Saisana, A. Saltelli, and S. Tarantola, “Uncertainty and s quality assessment of composite indicators,” *Journal of the* pp. 307–323, 2005.

21. A. Heiat, “Using an Excel extension for selecting the probab *in Education*, vol. 2, no. 1, pp. 95–100, 2005.

22. J. B. Schorling, J. Roach, M. Siegel, et al., “A trial of church African Americans,” *Preventive Medicine*, vol. 26, no. 1, pp

23. J. I. Cleeman, S. M. Grundy, D. Becker, et al., “Expert pane blood cholesterol in adults (adult treatment panel III),” *The* 285, no. 19, pp. 2486–2497, 2001.

24. J. T. DiPiro, R. L. Talbert, G. C. Yee, G. R. Matzke, B. G. Wel *Pathophysiologic Approach*, McGraw-Hill, New York, NY, USA

25. M. A. Koda-Kimble, L. Y. Young, W. A. Kradian, B. J. Gugliel *Therapeutics, The Clinical Use of Drugs*, Lippincott Williams