Journal of the Korean Data & Information Science Society
2011, **22**(3), 613–618

한국데이터정보과학회지

Cox proportional hazard model with L1 penalty[†]

Changha Hwang[1] · Jooyong Shim[2]

[1]Department of Statistics, Dankook University
[2]Department of Data Science, Inje University

Received 25 April 2011, revised 19 May 2011, accepted 23 May 2011

Abstract

The proposed method is based on a penalized log partial likelihood of Cox proportional hazard model with L1-penalty. We use the iteratively reweighted least squares procedure to solve L1 penalized log partial likelihood function of Cox proportional hazard model. It provide the efficient computation including variable selection and leads to the generalized cross validation function for the model selection. Experimental results are then presented to indicate the performance of the proposed procedure.

Keywords: Cox proportional hazard model, generalized cross validation function, iteratively reweighted least squares procedure, L1-penalty, least absolute shrinkage and selection operator.

1. Introduction

Let t_i be the response variables corresponding to covariate vector, $\boldsymbol{x}_i$ or transformation on it, where $i = 1, 2, \cdots, n$. In fact we can not observe t_i's but the observed variable, $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$, where $I(\cdot)$ denotes the indicator function and c_i is the censoring variable corresponding to $\boldsymbol{x}_i$ for $i = 1, 2, \cdots, n$. c_i's are assumed to be independently distributed with unknown survival distribution functions.
The proportional hazard model (Cox, 1972, 1975) includes the hazard function of the ith subject with covariate vector $\boldsymbol{x}_i$ of the form such that

$$h(t_i|\boldsymbol{x}_i) = h_0(t_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i), \tag{1.1}$$

where $h_0(t_i)$ is a unspecified baseline hazard function and $\boldsymbol{\beta}$ is a $p \times 1$ regression parameter vector. Generally, all the p covariates may not affect the survival patterns so that some β's may be zeros in true hazard function. Many variable selection techniques for linear regression models have been extended to the context of survival models, including the best-subset selection, stepwise selection, and Bootstrap procedures (Sauerbrei and Schumacher, 1992).

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028135).
[1] Department of Statistics, Dankook University, 126, Jukjeon-dong, Suji-gu, Yongin-si 448-701, Korea.
[2] Corresponding author: Department of Data Science, Inje University, Obang-Dong, Kimhae 621-749, Korea. E-mail: ds1631@hanmail.net

Recently the Lasso (least absolute shrinkage and selection operator, Tibshirani, 1997) has been proposed for Cox proportional hazards model. By shrinking some regression parameters to zero, this method provides the selection of important variables and the estimation of regression parameters simultaneously.

We consider the minimization of the penalized log partial likelihood function with L1 norm which is known to have the sparsity on estimation of regression parameters (Williams, 1995). We use the IRWLS (iteratively reweighted least squares) procedure to solve the penalized log partial likelihood function with L1 norm of Cox proportional hazard model. It provide the efficient computation including variable selection and leads to the generalized cross validation function for the model selection.

The rest of paper is organized as follows. In Section 2 we briefly review the penalized estimation for the Cox model. In Section 3 we propose IRWLS procedure to penalized estimation for the Cox model. In Section 4 we perform the numerical studies with simulated data sets. In Section 5 we give the conclusions.

2. Penalized estimation for the Cox model

Under the assumption of no ties, we can see that for each uncensored time y_i,

$$P(\text{ a failure in } [y_i, y_i + \triangle y)|R_i) \approx \sum_{j \in R_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_j) h_0(y_i) \triangle y, \tag{2.1}$$

$$P(\text{ a failure of i at } y_i| \text{ a failure in } R_i \text{ at } y_i) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_j)}, \tag{2.2}$$

where R_i is the risk set at time y_i. Cox (1972, 1975) proposed the proportional hazard model by treating the conditional likelihood (2.2) as an ordinary likelihood. The log partial likelihood of the Cox proportional hazard model is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}'\boldsymbol{x}_i - \log\{\sum_{j=i}^{n} \exp(\boldsymbol{\beta}'\boldsymbol{x}_j\} \right]. \tag{2.3}$$

When ties are present, the technique in Breslow (1974) can be used. The maximum likelihood estimate of $\boldsymbol{\beta}$ is obtained by solving $\dfrac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \mathbf{0}$, which usually needs the iterative methods. Under the proportional hazard model, the survival function is obtained as follows,

$$S(t : \boldsymbol{x}) = S_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}), \text{ where } S_0(t) = \exp(-\int_0^t h_0(u)du). \tag{2.4}$$

Breslow (1974) proposed the estimates of the survival function by assuming the piecewise constant baseline hazard functions. Tsiatis (1978) obtained the estimate of the survival function by assuming that the cumulative baseline hazard function is a step function.

To select important variables under Cox proportional hazard model, Tibshirani (1997) applied Lasso (Tibshirani, 1996) to Cox proportional hazard model:

$$min - l(\boldsymbol{\beta}), \text{ subject to } ||\boldsymbol{\beta}||_1 = \sum_i |\beta_i| < s, \tag{2.5}$$

where $s > 0$ is the specified parameter.

Zhang and Lu (2007) applied the adaptive Lasso to Cox proportional hazard model:

$$min - l(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_i|/|\widetilde{\beta}_i|, \tag{2.6}$$

where $\lambda > 0$ is the regularization parameter and $\widetilde{\boldsymbol{\beta}}$ is the estimator obtained by minimizing $-l(\boldsymbol{\beta})$.

3. IRWLS procedure to L1-penalized estimation for the Cox model

We consider the minimization of the penalized log partial likelihood function with L1 norm which is known to have the sparsity on estimation of $\boldsymbol{\beta}$ (Williams, 1995),

$$L(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1. \tag{3.1}$$

The penalized log partial likelihood function $L(\boldsymbol{\beta})$ in (3.1) is not differentiable with respect to $\boldsymbol{\beta}$, we need to modify $L(\boldsymbol{\beta})$ for IRWLS procedure.

We define the penalized log partial likelihood function given $\boldsymbol{\beta}^*$ as

$$L(\boldsymbol{\beta}|\boldsymbol{\beta^*}) = -l(\boldsymbol{\beta}) + \frac{\lambda}{2}\sum_{i=1}^{p}\left(\frac{\beta_i^2}{|\beta_i^*|} + |\beta_i^*|\right), \tag{3.2}$$

then $L(\boldsymbol{\beta}|\boldsymbol{\beta^*}) >= L(\boldsymbol{\beta})$ with equality if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ (Krishnapuram *et al.*, 2005) and $L(\boldsymbol{\beta}|\boldsymbol{\beta^*})$ is differentiable with respect to $\boldsymbol{\beta}$.

At l-th iteration of IRWLS procedure, we have

$$L(\boldsymbol{\beta}|\boldsymbol{\beta}^{(l)}) = -l(\boldsymbol{\beta}) + \frac{\lambda}{2}\sum_{i=1}^{p}\left(\frac{\beta_i^2}{|\beta_i^{(l)}|} + |\beta_i^{(l)}|\right). \tag{3.3}$$

We denote the gradient vector of $-l(\boldsymbol{\beta})$ by $G(\boldsymbol{\beta})$ and the Hessian matrix by $H(\boldsymbol{\beta})$. By second order Taylor series expansion, $-l(\boldsymbol{\beta})$ can be approximated as

$$-l(\boldsymbol{\beta}) \approx -l(\boldsymbol{\beta}^{(l)}) - \frac{1}{2}G(\boldsymbol{\beta}^{(l)})'H(\boldsymbol{\beta}^{(l)})^{-1}G(\boldsymbol{\beta}^{(l)}) + \frac{1}{2}||Y^{(l)} - X^{(l)}\boldsymbol{\beta}||^2, \tag{3.4}$$

where $X^{(l)}$ is obtained from the Cholesky decomposition of $H(\boldsymbol{\beta}^{(l)})$ such as $H(\boldsymbol{\beta}^{(l)}) = X^{(l)'}X^{(l)}$ and $Y^{(l)} = (X^{(l)'})^{-1}(H(\boldsymbol{\beta}^{(l)})\boldsymbol{\beta}^{(l)} - G(\boldsymbol{\beta}^{(l)}))$. The penalized log partial likelihood function (3.3) can be written as

$$L(\boldsymbol{\beta}|\boldsymbol{\beta}^{(l)}) = \frac{1}{2}||Y^{(l)} - X^{(l)}\boldsymbol{\beta}||^2 + \frac{\lambda}{2}\sum_{i=1}^{p}\left(\frac{\beta_i^2}{|\beta_i^{(l)}|} + |\beta_i^{(l)}|\right), \tag{3.5}$$

which is solved through the IRWLS procedure.

At $(l+1)$th iteration, $\boldsymbol{\beta}^{(l+1)}$ is obtained by minimizing $L(\boldsymbol{\beta}|\boldsymbol{\beta}^{(l)})$ with respect to $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^{(l+1)} = (X^{(l)\prime}X^{(l)} + \lambda W^{(l)})^{-1}X^{(l)\prime}Y^{(l)} \tag{3.6}$$

where $W^{(l)}$ is the diagonal matrix consisted of $1/|\beta_i^{(l)}|, i = 1, \cdots, p$.

During iteration, we find that some β_i's tend to zero keeping the value of objective function $L(\boldsymbol{\beta})$ decreasing. This motivates that we can find sparse estimates of $\boldsymbol{\beta}$ which provides decreasing value of the penalized log partial likelihood function $L(\boldsymbol{\beta})$ at the same time.

Algorithm of L1-penalized estimation for Cox proportional hazard model:

1. Set $v = (1:p)'$ and $\boldsymbol{\beta}(v)^{(0)}$.
2. Find solution $\boldsymbol{\beta}(v)^{(l+1)}$ which minimizes $L(\boldsymbol{\beta}(v)|\boldsymbol{\beta}(v)^{(l)})$.
3. Set $\beta_i^{(l+1)} = 0$ which is very close to zero. Find $v = \{i|\beta_i^{(l+1)} \neq 0\}$.
4. Iterate 2-4 until $|L(\boldsymbol{\beta}(v)|\boldsymbol{\beta}(v)^{(l+1)}) - L(\boldsymbol{\beta}(v)|\boldsymbol{\beta}(v)^{(l)})| <$ Tol.

The functional structures of L1-penalized log partial likelihood for Cox proportional hazard model is characterized by the regularization parameter λ.

With the final estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}} = (X'X + \lambda W)^{-1}X'Y$, the ridge regression estimator, the number of effective parameters (Moody, 1992) is approximated by

$$d(\lambda) = tr\{X(X'X + \lambda W)^{-1}X'\}. \tag{3.7}$$

Thus, we have the generalized cross validation (GCV) function (Craven and Wahba, 1979, Tibshirani, 1997) as follows:

$$GCV(\lambda) = \frac{1}{n}\frac{-l(\widehat{\boldsymbol{\beta}})}{(1 - d(\lambda)/n)^2}, \tag{3.8}$$

which is different from GCV function used in kernel regression of Cho *et al.* (2010), Hwang and Shim (2010), Shim (2005), Shim and Lee (2009).

4. Numerical studies

We illustrate the performance of the IRWLS procedure to penalized estimation for Cox proportional hazard model by comparing its performance with the Cox method with adaptive LASSO (2.6) and the Cox method (2.3) via 50 simulated data sets. In each data set of size $n = 44$, the hazard function of i th subject is set to $h(t_i|\boldsymbol{x}_i) = 0.1\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ with $\boldsymbol{\beta} = (1, 0, 1, 0, 0, 0)'$. The covariate $\boldsymbol{x}_i$ is generated from $N(\mathbf{0}_{6\times 1}, \mathrm{I}_6 \times 6)$, the survival time t_i is set to $t_i = -\log(u_i)/h(t_i|\boldsymbol{x}_i)$ with u_i generated from $U(0,1)$ and the censored time c_i is generated from $U(0,6)$ and the observed time y_i is $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$, $i = 1, 2, \cdots, 44$. From 50 data sets we obtained mean squared error of ($\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}$) and its

standard error for each method. As results we obtained mean squared error and its standard error for the proposed method as (0.0636, 0.0161), (0.1406, 0.0188) for the Cox method with adaptive LASSO and (0.4578, 0.1677) for the Cox method. The box plots of each estimated regression parameter for each method are shown as in Figure 4.1. From the results we can see that the proposed method provides a little lower estimates for β_1 and β_3 whose true values are 1 than other methods, but generally the proposed method shows a more satisfying results (stable estimates) for given simulated data sets than other methods.

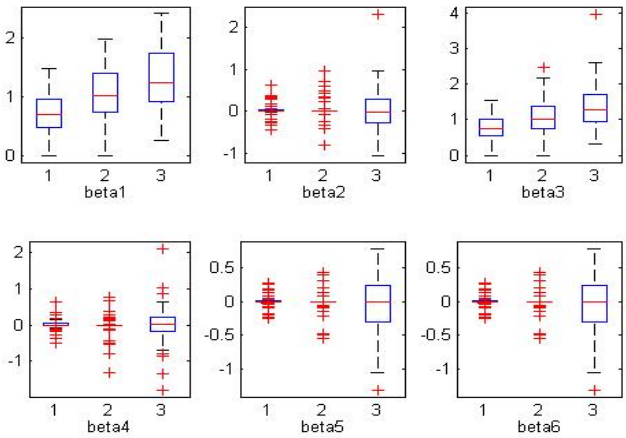

Figure 4.1 Box plots of estimated regression parameters
1: proposed, 2: the Cox method with adaptive LASSO, 3: the Cox method

5. Conclusions

In this paper, we proposed the efficient method to solve the penalized log partial likelihood function with L1 norm of Cox proportional hazard model. We use IRWLS procedure to solve L1 penalized log partial likelihood function, which provides the efficient estimation and variable selection simultaneously. And it leads to the generalized cross validation function for the model selection. From the simulated data we found that the proposed method provides the stable results for variable selection.

References

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.

Cho, D. H., Shim, J. and Seok, K. H. (2010). Doubly penalized kernel method for heteroscedastic autoregressive data. *Journal of the Korean Data & Information Science Society*, **21**, 155-162.

Cox, D. R. (1972) Regression models and life tables (with discussions). *Journal of the Royal Statistical Society,* Series B, **7**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerical Mathematics*, **31**, 377-403.

Hwang, C. and Shim, J. (2010). Semiparametric support vector machine for accelerated failure time model. *Journal of the Korean Data & Information Science Society*, **21**, 467-477.

Krishnapuram, B., Carlin, L., Figueiredo, M. A. T. and Hartermink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 957-968.

Moody, J. E. (1992). The effective number of parameters: An analysis of generalisation and regularisation in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4*, edited by J. E. Moody, S. J. Hanson and R. P. Lippmann, 847-854.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistical Medicine,* **11**, 2093-2099.

Shim, J. (2005). Censored kernel ridge regression. *Journal of the Korean Data & Information Science Society*, **16**, 1045-1052.

Shim, J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of the Korean Data & Information Science Society*, **20**, 467-4720.

Tsiatis, R. (1978). *A heuristic estimate of the asymptotic variance of survival probability in Cox' regression model*, Technical report of University of Wisconsin, number 524.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, **7**, 117-143.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691-703.