Statistica Sinica Preprint No: SS-2016-0401R1	
Title	Asymptotic Behavior of Cox's Partial Likelihood and its Application to Variable Selection
Manuscript ID	SS-2016-0401
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0401
Complete List of Authors	Runze Li Jian-Jian Ren Guangren Yang and Ye Yu
Corresponding Author	Runze Li
E-mail	rzli@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

probability as the sample size tends to infinity.

With the aid of the asymptotic property of partial likelihood, we study the selection of regularization parameter in penalized partial likelihood for variable selection. Tibshirani (1997) proposed penalized partial likelihood with LASSO penalty for the Cox model. Fan and Li (2002) proposed the partial likelihood with the SCAD penalty for the Cox model and showed that under certain regularity conditions, the resulting estimate enjoys the oracle property. Zhang and Lu (2007) and Zou (2008) further proposed adaptive LASSO for the Cox model to improve the SCAD procedure in terms of computational efficiency, while retaining the oracle property. The oracle property depends on the choice of the regularization parameter in penalized partial likelihood. It is well known that the regularization parameter controls the model complexity of the selected model and plays a crucial role in these variable selection procedures. The issue of regularization parameter selection for penalized partial likelihood has not been systematically studied, in part because the asymptotic behavior of partial likelihood was not well understood. Wang, Li, and Tsai (2007) studied the selection of regularization parameter in the penalized least squares for linear regression models. They showed that with a positive probability, the *generalized cross-validation* (GCV, Craven and Wahba (1979)) selector yields an over-fitted model, and therefore this procedure does not enjoy the oracle property.

In this paper we prove that the GCV selector for the SCAD method for the Cox model enjoys model selection consistency, in contrast to its model selection inconsistency in the least squares setting as demonstrated in Wang, Li, and Tsai (2007). Although GCV is equivalent to AIC and the C_p in the context of linear regression models, AIC and C_p yield an overfitted models with a positive probability, and thus are not model selection consistent.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

The rest of this paper is organized as follows. Section 2 studies the asymptotic behavior of the partial likelihood of the Cox model. We study the regularization parameter selection for the penalized partial likelihood in Section 3. Simulation study and a data example are presented in Section 4. Proofs are given in the Appendix.

2. Asymptotic Behavior of Cox's Partial Likelihood

Let T and $\mathbf{X} = (X_1, \cdots, X_d)^T$ be the survival time and associated d-dimensional vector of covariates, respectively. Consider the Cox proportional hazard regression model:

$$h(t \,|\, \mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}), \tag{2.1}$$

where $\boldsymbol{\beta}$ is the regression coefficient vector, and $h(t \,|\, \mathbf{x})$ is the conditional hazard function of T given $\mathbf{X} = \mathbf{x}$ with $h_0(t)$ as an arbitrary baseline hazard function. Suppose that $(T_1, \mathbf{x}_1), \ldots, (T_n, \mathbf{x}_n)$ is a random sample of $(T, \mathbf{X})$, and the observed right censored survival data are as follows: $(V_1, \delta_1, \mathbf{x}_1)$, $\ldots$, $(V_n, \delta_n, \mathbf{x}_n)$, where $V_i = \min\{T_i, C_i\}, \delta_i = I\{T_i \leq C_i\}$, and C_i is the *right censoring variable* independent of T_i given $\mathbf{X} = \mathbf{x}_i$. Without loss of the generality, assume that there are no ties among observed continuous random variables V_i's. The log-partial likelihood function of the observed data is

$$\ell_c(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log \Big(\textstyle\sum_{j=1}^{n} I\{V_j \geq V_i\} \exp(\mathbf{x}_j^T \boldsymbol{\beta}) \Big). \tag{2.2}$$

(Cox (1975)). The goal is to study the asymptotic behavior of $\ell_c(\boldsymbol{\beta})$. We first illustrate the different behaviors of the log-partial likelihood and the likelihood of an i.i.d. sample by an example.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

Example 1. Suppose that we have an i.i.d. random sample $\{Y_1, \ldots, Y_n\}$ from a population with probability density/mass function $f(y;\theta)$, so $\ell(\theta) = \sum_{i=1}^{n} \log\{f(Y_i;\theta)\}$ is the log-likelihood function. By the weak law of large number, $n^{-1}\ell(\theta) \to E\log\{f(Y;\theta)\}$ in probability under mild regularity conditions. Furthermore, under mild regularity conditions, the maximum partial likelihood estimator, the maximizer of $\ell_c(\boldsymbol{\beta})$, behaves the same as the ordinary maximum likelihood estimator, the maximizer of $\ell(\theta)$, in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency. See, for example, Murphy and van der Vaart (2000). Here, we numerically illustrate that

$$n^{-1}\ell_c(\boldsymbol{\beta}) \to \infty \quad \text{as} \quad n \to \infty. \tag{2.3}$$

We generated a random sample of size n from the proportional hazard model

$$h(t|x) = h_0(t)\exp(X\beta),$$

where $h_0(t) \equiv 1$, $\beta = 1$ and $X \sim N(0,1)$. The censoring variable C was generated from an exponential distribution with mean U. Therefore, the average censoring rate varies with different values of U. We list several values of U in Table 2.1 together with their corresponding average censoring rates, $1 - E\,I(T \le C) \widehat{=} 1 - \rho_1$, and take 10 different values of n ranging from $4(= 2^2)$ to $1024(= 2^{10})$. Figure 2.1 depicts the scatter plot of $\log(n)$ versus $-n^{-1}\ell_c$ based on a set of typical samples based on the different U listed in Table 2.1. Figure 2.1 clearly suggests that $-n^{-1}\ell_c$ increases at $\log(n)$ rate.

We next show that $-n^{-1}\ell_c(\boldsymbol{\beta})$ tends to infinite at the rate of $\log(n)$ using techniques

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0401

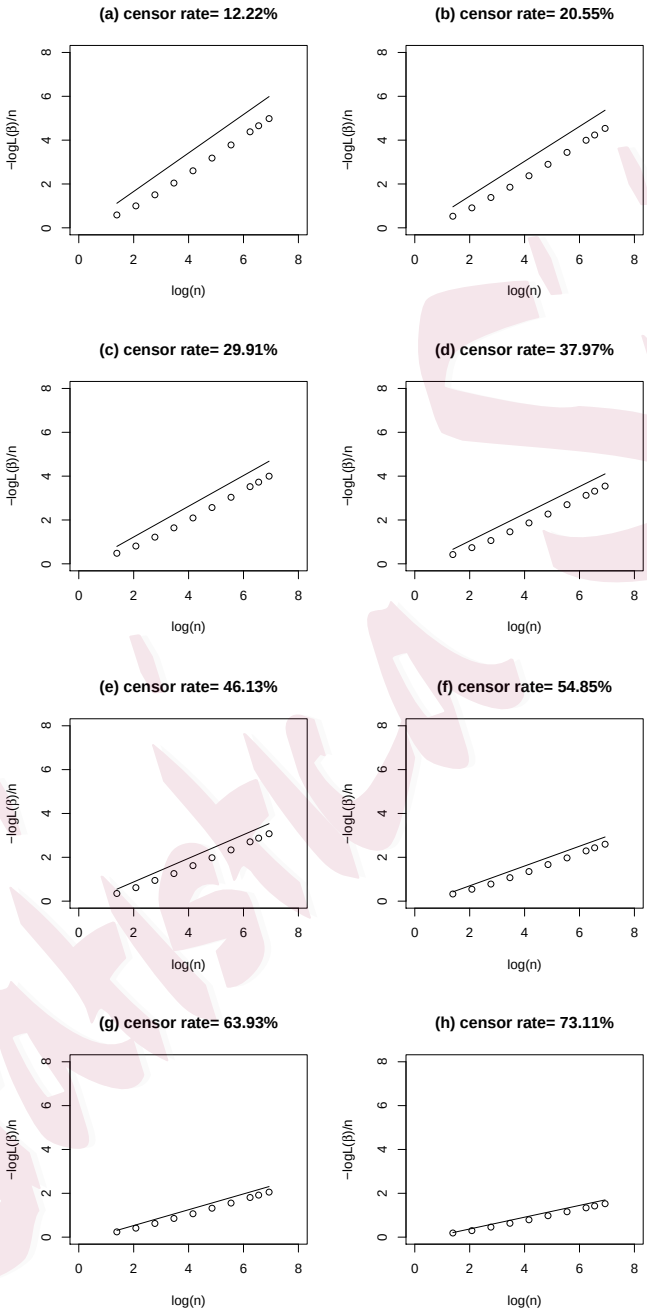

Figure 2.1: Plot of $\log(n)$ versus $-n^{-1}\ell_c$. 'o' is the scatter plot of $\log(n)$ versus $-n^{-1}\ell_c$ based on a typical simulated data set. The solid line in each plot is $\log(n)\widehat{\rho}_1 - \boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_0$ with $\boldsymbol{\beta} = 1$, where $\widehat{\rho}_1$ is an estimate of $EI\{T \le C\}$ and $\widehat{\boldsymbol{\mu}}_0$ is an estimate of $E\{I\{T \le C\}X\}$.