Scientific Research Publishing

Modeling and Numerical Simulation of Material Science, 2021, 11, 47-53
https://www.scirp.org/journal/mnsms
ISSN Online: 2164-5353
ISSN Print: 2164-5345

Sequential Shrinkage Estimate for COX Regression Models with Uncertain Number of Effective Variables

Haibo Lu, Juling Zhou, Cuiling Dong

School of Mathematical Sciences, Xinjiang Normal University, Urumqi, China
Email: andyluhaibo@foxmail.com

How to cite this paper: H.B., Lu, Zhou, J.L. and Dong, C.L. (2021) Sequential Shrinkage Estimate for COX Regression Models with Uncertain Number of Effective Variables. *Modeling and Numerical Simulation of Material Science*, **11**, 47-53.
https://doi.org/10.4236/mnsms.2021.113004

Received: June 17, 2021
Accepted: July 19, 2021
Published: July 22, 2021

Abstract

In the applications of COX regression models, we always encounter data sets that contain too many variables that only a few of them contribute to the model. Therefore, it will waste much more samples to estimate the "noneffective" variables in the inference. In this paper, we use a sequential procedure for constructing the fixed size confidence set for the "effective" parameters to the model based on an adaptive shrinkage estimate such that the "effective" coefficients can be efficiently identified with the minimum sample size. Fixed design is considered for numerical simulation. The strong consistency, asymptotic distributions and convergence rates of estimates under the fixed design are obtained. In addition, the sequential procedure is shown to be asymptotically optimal in the sense of Chow and Robbins (1965).

Keywords

Sequential Estimate, COX Regression Model, Stopping Time, Minimum Sample Size

1. Introduction

The COX proportional hazards model is a popular choice for the analysis of censored survival data with covariates, illustrated in [1] [2] [3]. It has been widely used in many areas, such as biomedical research and engineering, for assessing covariate effects on the time to some events in the presence. However, in applications such as Biology, Engineering and Epidemiology there are data sets that usually have a large number of explanatory variables but only a few of them contributes to the model. They were called effective variables in [4]. Many methods are focused on how to identify the effective variables such as LASSO and

LARS, see in [5] and [6], however, people also want to know how many samples can identify the effective variables and simultaneously make the parameter estimates achieve a pre-specified accuracy. It is very important to those who care about the cost of samples such as Biology and Epidemiology. For linear regression model, Wang and Chang propose a sequential shrinkage estimate method to identify the effective variables and attain accuracy of parameter estimate in [4]. For COX regression models, similar methods have not been proposed, so there is still a lot of work to do for this problem.

For handling the problem mentioned above, we propose a sequential procedure for constructing the fixed size confidence set for effective parameters based on an adaptive shrinkage estimate (ASE) such that the effective coefficients can be efficiently identified with the minimum sample size. Suppose the conditional hazard rate of a survival time, T, given the regressor vector, X, is written as

$$h(t \mid x) = h_0(x)\exp(\beta' X), t \geq 0 \tag{1}$$

In the paper, it will be studied under fixed design and the consistency and asymptotic properties of the proposed estimator will be obtained under this design. The rest of this paper is organized as follows. In Section 2, we will give the adaptive shrinkage estimate (ASE) based on the Maximum Partial Likelihood Estimate (MPLE) of COX regression models and their asymptotic properties. In section 3, sequential sampling strategy based on ASE and stopping rule as well as random size confident set is presented. In Section 4, an example with numerical simulation is given to illustrate the performance of the proposed method via sequential fixed size confidence estimation using synthesized data sets.

2. Sequential Adaptive Shrinkage Estimate

2.1. Asymptotic Properties of MPLE

Let T_i and C_i be the potential failure time and censoring time of the i-th $(i \in N_+)$ subject from a random sample with n individuals, respectively, and $X_i = (X_{i1}, X_{i2}, \cdots, X_{in})^{\mathrm{T}}$ be a p-dimensional vector of covariates which assumed to be time-independent throughout this paper for the i-th individual. Assume that T_i and C_i are conditionally independent given X_i. In practice, the failure time T_i might not always be observed due to censoring because of the termination of study or early withdrawal from the study. What we can actually observe are $Y_i = \min\{T_i, C_i\}$, the smaller of the failure time and the censoring time, and $\delta_i = I\{T_i \leq C_i\}$, the indicator that failure has been observed. The data then consist of the triplets $(Y_i, \delta_i, X_i), i = 1, 2, \cdots, n$. Suppose there is no tie among failure times. Let $t_1, t_2, \cdots, t_n$ denote the N ordered times of observed failures and (j) be the label of the individual that fails at t_j. Let R_j be the risk set at time t_j, *i.e.* $R_j = \{i : Y_i \geq t_j\}$. The partial likelihood of the model (1) is defined as

$$\prod_{j=1}^{N} \frac{\exp\left(\beta^{\mathrm{T}} X_{(j)}\right)}{\sum_{i \in R_j} \exp\left(\beta^{\mathrm{T}} X_{(i)}\right)} \tag{2}$$

and the log partial likelihood is then,

$$L(\beta)=\sum_{j=1}^{N}\left\{\beta^{\mathrm{T}}X_{(j)}-\log\left[\sum_{i\in R_j}\exp\left(\beta^{\mathrm{T}}X_i\right)\right]\right\} \tag{3}$$

The maximum partial likelihood estimate of β, $\tilde{\beta}$, is found by solving the score equation $U(\beta)=0$, where

$$U(\beta)=\frac{\partial L(\beta)}{\partial\beta}=\sum_{j=1}^{N}\left\{X_{(j)}-\frac{\sum_{i\in R_j}X_i\exp\left(\beta^{\mathrm{T}}X_i\right)}{\sum_{i\in R_j}\exp\left(\beta^{\mathrm{T}}X_i\right)}\right\} \tag{4}$$

2.2. Adaptive Shrinkage Estimate

Let $\kappa=\kappa(n)$ be a non-random function of n such that for some $0<\delta<1/2$ and $\gamma>0$, $n^{1/2}\kappa\to 0$ and $n^{1/2+\gamma\delta}\kappa\to\infty$, as $n\to\infty$. In this paper, we need the following assumptions:

(A1) x_i satisfies $\sup_i\|x_i\|<\infty$, and the residual term

$$\varepsilon_i=\hat{\Lambda}(Y_i)\exp\left(\hat{\beta}^{\mathrm{T}}X_i\right)$$

has $E|\varepsilon_i|^{\zeta}<\infty$ for some $\zeta>2$, where $\hat{\Lambda}$ is some cumulative baseline function.

(A2) $\lim_{n\to\infty}I_n(\beta)/n=\Sigma$, where $I_n(\beta)$ is the information matrix of β and Σ is a positive matrix.

Then, Theorem 3.1 in [7] implies that $n^{1/2-\eta}\left(\tilde{\beta}_n-\beta_0\right)=O(1)$ almost surely as n tends to ∞ for some $\eta>0$. Define

$$\kappa_{nj}=\kappa\left|\tilde{\beta}_{nj}\right|^{-\gamma}$$

with $\tilde{\beta}_{nj}$ being the j-th components of $\tilde{\beta}_n$. From (16) and asymptotic property of $\tilde{\beta}_n$ we have $n^{1/2}\kappa_{nj}\to 0\times I\left(\beta_{0j}\neq 0\right)+\infty\times I\left(\beta_{0j}=0\right)$ almost surely as $n\to\infty$. Where $I(\cdot)$ denotes the indicator function and we presume that $0\times\infty=0$. Similar to Wang and Chang, define $\hat{\beta}_n=I_n(\varepsilon)\tilde{\beta}_n$ as an adaptive shrinkage estimate (ASE) of β_0, where $I_n(\varepsilon)=diag\left\{I_{n1}(\varepsilon),I_{n2}(\varepsilon),\cdots,I_{np}(\varepsilon)\right\}$ is a $p\times p$ diagonal matrix. So far, we get good statistical properties of the proposed ASE estimate under non-random sample size, but our goal is to determine a sample size under which the ASE attains the required accuracy. To this end, we will introduce the sequential sampling scheme based on the ASE below. It is known that construction of the confidence set for β_0 depends on the asymptotic distribution of $\hat{\beta}_n$ and sample size under sequential analysis is a random variable. So we need to study asymptotic properties of ASE under random sample size. Fortunately, property of uniform continuity in probability, see in [8] and [9], is a sufficient condition such that the randomly stopped sequence has the same asymptotic distribution as the fixed sample size estimate. That is, $\sqrt{n}\left(\hat{\beta}_n-\beta_0\right),n=1,2,\cdots$, has the property of uniform continuity in probability, which indicates the following Theorem holds.

Theorem 1. Suppose that the (A1) and (A2) are satisfied, and let $N(t)$ be a

positive integer-valued random variable such that $N(t)/t$ converges to 1 in probability as $t \to \infty$. Then

$$\sqrt{N(t)}\left(\hat{\beta}_{N(t)} - \beta_0\right) \to N\left(0, I_0 \Sigma I_0^{-1}\right)$$

in distribution as $t \to \infty$.

From Theorem 1, we can construct a confidence set of β_0 and a stopping rule on sequential sampling procedure to determine final sample size. Let $\{(y_i, x_i) : i = 1, 2, \cdots, k\}$ be the first k observations and denoted by C_k. Define a stopping rule N_d as

$$N = N_d \equiv \inf\left\{k : \frac{d^2}{a_k^2} \geq \nu_k, \forall k \geq n_0\right\} \tag{5}$$

For sequential estimation procedure, one new observation is collected at a time until the stopping criterion is satisfied. When the stopping rule holds, based on N samples a confidence set of β_0 is constructed as follow,

$$R_N = \left\{Z \in R^p : \frac{S_N}{N} \leq \frac{d^2}{\nu_N}; I_{N_j}(\varepsilon) = 0 \to z_j = 0, 1 \leq j \leq p\right\} \tag{6}$$

where $S_N = \left(Z_{N_1} - \hat{\beta}_{N_1}\right)^{\mathrm{T}} \tilde{\Sigma}_{11} \left(Z_{N_1} - \hat{\beta}_{N_1}\right)$. Properties of the sequential procedure and the confidence set R_N are summarized below.

Theorem 2. Assume that the (A1) and (A2) are satisfied, and let N be the stopping time defined in Equation (5). Then 1) $\lim_{d \to 0} d^2 N / a^2 \nu = 1$ almost surely; 2) $\lim_{d \to 0} d^2 N / a^2 \nu = 1$; 3) $\lim_{d \to 0} d^2 E(N) / a^2 \nu = 1$; 4) $\lim_{d \to 0} \hat{p}_0(N) = p_0$ almost surely; 5) $\lim_{d \to 0} E\left(\hat{p}_0(N)\right) = p_0$ where ν is the maximum eigen-value of matrix $I_0 \Sigma^{-1} I_0$.

3. Example and Simulation

We evaluate the performance of the proposed method via sequential fixed size confidence estimation using synthesized data sets. As mentioned previously, by the definition of the stopping rule, when sampling is stopped, the final confidence ellipsoid constructed will have the prescribed precision and coverage probability. Thus, we can compare the average stopping times of procedures based on MPLE and ASE. Since the proposed method ignores the non-effective variables, we expect the average stopping time to be significantly smaller than that of the procedure based on MPLE with no variable identification mechanism. If the p_0 variables are known in advance, then the most efficient procedureis, of course, to use only these p_0 variables. Therefore, we also construct a sequential procedure under such a situation, and the results of the cases with known p_0 can serve as the baseline, in which the smallest sample size is achieved, asymptotically.

The synthesized data sets for the model with fixed designs are generated as follows: the regressor x_i are generated independently from a standard multivariate normal distribution with mean 0 and identity covariance matrix beforehand, and the error term e_i is independently drawn from the standard normal distribution for each $i \geq 1$. The system error is assumed to follow the standard

normal distribution. The response generated by model (1) with the arbitrary $h_0(t)=t^2$ without loss of the generality and the true parameter $\beta_0=(-1.2,2.0,0,0,0,0,0,0,0,0)$ with 8 non-effective variables. Different precisions of confidence ellipsoid $d\in\{0.3,0.4,0.5,0.6\}$ are chosed with coverage probability equal to 95% $\alpha=0.05$ in the simulation. We choose $\gamma=1$, $\delta=0.45$ and $\theta=0.75$ in analyzing simulated data. When applying the ASE method, the regularization parameter ε needs to be determined by some model selection criteria, as the AIC, BIC together with a GCV method. For convenience, we only use BIC to illustrate our method,

$$\mathrm{BIC}=-2\left(\sum_{j=1}^{N}\left(\beta^{\mathrm{T}}X_{(j)}-\log\left(\sum_{j\in R_j}\exp\left(\beta^{\mathrm{T}}X_i\right)\right)\right)\right)+\log(n)\times df/n,$$

where df is the number of the non-zero components in β.

Table 1 state results of sequential sampling method for COX regression. In the table, we list final sample size N (stopping time), $\kappa=d^2N/a^2\nu$ and empirical coverage probability CP of the 95% confidence set R_N. For all of the three cases: MPLE, MPLE_{p_0}, ASE, the value κ of is very close to 1, and the empirical coverage probability CP approaches the Normal 95% as d decreases, as stated in Theorem 2. However, the sample size N of MPLE are much larger than those of the other two cases, and ASE has sample size very close to those of MPLE_{p_0}. In conclusion, the proposed ASE is more efficient than MPLE.

Table 2 reports powers of identity effective variables and effective variables and estimates of the regression coefficients for COX regression. We can see that numbers of incorrectly identified zero variables (N_{ic}^*) using ASE is almost close to 0, and the number of correctly identified zero variables (N_c^*) are all very close to the true number of effective variables (2 and 8). These results suggest that $\hat{p}_0$ is a good estimator of p_0 under the sequential sampling method based on ASE. The MPLE procedure does not identify the effective variables, so N_c^* and N_{ic}^* are not available. In addition, all of parameter estimates of effective variables are very close to the true values.

Table 1. Results of sequential sampling method based on ASE, MPLE with all variables and MPLE_{p_0} with only p_0 non-zero variables for COX regression model.

		$\beta_0=(-1.2,2.0,0,0,0,0,0,0,0,0)$								
		MPLE_{p_0}			ASE			MPLE		
Design	d	N	κ^*	CP	N	κ	CP	N	κ	CP
fixed	0.6	95.740 (14.75)**	1.028	0.95	107.62 (17.40)	1.044	0.93	305.8 (23.194)	1.01	0.94
	0.5	130.44 (19.75)	1.017	0.98	141.52 (19.13)	1.034	0.93	419.84 (29.586)	1.006	1
	0.4	198.18 (25.587)	1.008	0.93	199.394 (25.311)	1.021	0.928	632.936 (37.868)	1.003	0.93
	0.3	359.68 (38.211)	1.004	0.95	333.676 (37.087)	1.017	0.94	1100.05 (44.707)	1.002	0.97

$\kappa^*=d^2N/(a^2\nu)$; CP^+ is the empirical coverage probability of 95% confidence ellipsoid region R_N; **Empirical standard deviations are in parentheses.

Table 2. Power of variable identification and estimation of nonzero components under sequential sampling method based on ASE and MPLE with COX regression model.

		$\beta_1 = -1.2, \beta_2 = 2.0$							
		ASE				MPLE			
Design	d	N_{ic}^*	N_c^*	β_1	β_2	N_{ic}^*	N_c^*	β_1	β_2
fixed	0.6	0	7.876	−1.263 (0.155)	2.10 (0.210)	-	-	−1.220 (0.09)	2.051 (0.01)
	0.5	0	7.916	−1.243 (0.129)	2.072 (0.173)	-	-	−1.214 (0.007)	2.041 (0.096)
	0.4	0	7.932	−1.232 (0.105)	2.053 (0.142)	-	-	−1.214 (0.064)	2.011 (0.082)
	0.3	0	7.956	−1.220 (0.076)	2.037 (0.107)	-	-	−1.202 (0.042)	2.005 (0.057)

N_{ic}^* and N_c^* are the average number of zero components in β correctly identified and nonzero components incorrectly estimated as zero values, respectively.

4. Conclusion

Based on an ASE estimate of the parameter in COX regression model, a sequential sampling procedure is constructed to estimate the minimum sample size to identify the effective variables and simultaneously make estimate of parameters with required accuracy. We prove that the proposed sequential procedure is asymptotically optimal in the sense of Chow and Robbins [10]. Simulation studies show that the proposed method can save a large sample size compared to the traditional sequential sampling method. However, this paper supposes the dimension of variables is fixed, not varying as sample size. Our future work is to investigate the properties of sequential sampling method with varying number of variables as sample size.

Supported

This research was supported by Research projects of universities in Xinjiang Uygur Autonomous Region under Grant No. XJEDU2016I033 and Xinjiang Normal University postdoctoral research foundation under Grant No. XJNUBS1539.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] COX, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society*: *Series B*, **34**, 187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[2] COX, D.R. (1975) Partial Likelihood. *Biometrika*, **62**, 269-276. https://doi.org/10.1093/biomet/62.2.269

[3] Andersen, P.K. and Gill, R.D. (1982) COX's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, **10**, 1100-1120. https://doi.org/10.1214/aos/1176345976

[4] Wang, Z.F. and Chang, Y.I. (2013) Sequential Estimate for Linear Regression Mod-

els with Uncertain Number of Effective Variables. *Metrika*, **76**, 949-978. https://doi.org/10.1007/s00184-012-0426-4

[5] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[6] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Journal of Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[7] Tsiatis, A.A. (1981) A Large Sample Study of COX's Regression Model. *Annals of Statistics*, **9**, 93-108. https://doi.org/10.1214/aos/1176345335

[8] Anscombe, F.J. (1952) Large Sample Theory of Sequential Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **48**, 600-607. https://doi.org/10.1017/S0305004100076386

[9] Woodroofe, M. (1982) Nonlinear Renewal Theory in Sequential Analysis. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611970302

[10] Chow, Y.S. and Robbins, H. (1965) On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. *Journal of Annals of Mathematical Statistics*, **36**, 457-462. https://doi.org/10.1214/aoms/1177700156