Dhaka Univ. J. Sci. 67(2): 111- 116, 2019 (July)

Model Selection Strategy for Cox Proportional Hazards Model

Fabiha Binte Farooq and Md. Jamil Hasan Karami*

Department of Statistics, Dhaka University, Dhaka- 1000, Bangladesh

(Received: 20 February 2019; Accepted: 23 June 2019)

Abstract

Often in survival regression modelling, not all predictors are relevant to the outcome variable. Discarding such irrelevant variables is very crucial in model selection. In this research, under Cox Proportional Hazards (PH) model we study different model selection criteria including Stepwise selection, Least Absolute Shrinkage and Selection Operator (LASSO), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and the extended versions of AIC and BIC to the Cox model. The simulation study shows that varying censoring proportions and correlation coefficients among the covariates have great impact on the performances of the criteria to identify a true model. In the presence of high correlation among the covariates, the success rate for identifying the true model is higher for LASSO compared to other criteria. The extended version of BIC always shows better result than the traditional BIC. We have also applied these techniques to real world data.

Keywords: AIC, BIC, LASSO, Stepwise regression.

I. Introduction

Survival analysis is a special branch of statistics dealing with statistical methods for analyzing survival data available from clinical trials and biomedical studies. An important part of survival analysis is to fit a model based on the relationship between response variables and covariates. One way to achieve this is to search for a theoretical model that fits the observed data and identify the most important factors. Difficulties arise due to the presence of censored observations in survival analysis. So, it has been always a tricky task to select the most important covariates in survival regression. Many model selection techniques have been suggested in the literature. Among them the most widely used techniques are Stepwise selection (Efromyson),[5] Akaike Information Criterion (AIC) (Akaike),[1] Improved AIC (Hurvich and Tsai),[7] Delta AIC, Bayesian Information Criterion (BIC) (Schwarz),[12] Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani).[9] AIC has a tendency to pick the larger models. BIC tends to pick the smaller ones. On the other hand, LASSO maintains a balance between the larger and smaller models. In this research, with Cox model, we aim to explore several model selection criteria including Stepwise selection, LASSO, AIC, BIC and the extended versions of AIC and BIC. We also intend to evaluate their performances with simulated data as well as real world data.

The rest of the paper is organized as follows. In Section **II**, we briefly discuss the methods used in this research. Section **III**, shows the results of simulation study along with real life example. Section **IV** contains a conclusion and includes further research of this study.

II. Methods

Model and Estimation

The basic model for survival data is the proportional hazards model. Suppose that a number of patients are given randomly either a standard treatment or a new treatment, and let $h_S(t)$ and $h_N(t)$ be the hazards of death at time t for patients on the standard treatment and new treatment respectively. The simple proportional hazards model can be expressed in the form (Collett),[3] $h_N(t) = \psi\, h_S(t)$, for any non-negative value of t, where ψ is a constant and known as hazard ratio. Suppose that the hazard function for the i^{th} individual is $h_i(t)$, $i = 1, 2, ..., n$. Also let $h_0(t)$ be the hazard function for an individual on the standard treatment (baseline hazard). Then the hazard function for an individual on the new treatment is $\psi\, h_0(t)$. The hazard ratio, ψ cannot be negative and according to Cox,[4] it is convenient to set $\psi = exp(\boldsymbol{\eta}_i)$, where $\boldsymbol{\eta}_i$ is a linear combination of the p covariates, i.e. $\boldsymbol{\eta_i} = \sum_{j=1}^{p} \beta_j x_{ij}$, where β_j is the corresponding regression coefficient. Then the general proportional hazards model becomes

$$h_i(t) = h_0(t)\, exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}),\ \ i = 1, 2, ..., n \quad (1)$$

For this semiparametric model in equation (1), the relevant likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta' x}_{(j)})}{\sum_{l \in R(t(j))} \exp(\boldsymbol{\beta' x}_l)}, \quad (2)$$

where $\boldsymbol{x}_{(j)}$ is the vector of covariates for the individual who dies at the j^{th} ordered time, $t_{(j)}$. The denominator represents the sum of the values of $exp(\boldsymbol{\beta' x})$ over all individuals who are at risk at time $t_{(j)}$. Using equation (2), the regression parameter $\boldsymbol{\beta'}$ can be estimated by Newton-Raphson iterative procedure.

Model Selection Criteria

With the Cox model, we study several model selection criteria. Firstly, we discuss Akaike Information Criterion (AIC) (Akaike)[1] and it has the following form

$$AIC = -2log(L(\hat{\boldsymbol{\beta}})) + 2p,$$

where the first term consists of negative of log likelihood and the second term is a penalty, which is twice the number of parameters in the model. Hurvich, Simonoff and Tsai[7] show that in non-parametric regression model the AIC selects model with excess number of covariates. They suggested a corrected version of AIC where the penalty

*Author for correspondence. e-mail: karami.stat@du.ac.bd

term is replaced by $\frac{n(p+1)}{n-(p+2)}$, where n is the total number of observations. Again for Cox model, Therneau and Grambsch[8] suggested that in the penalty term, n should be replaced by the number of uncensored events, r. Hence, the corrected AIC can be written as

$$AICc = -2\log\left(L(\hat{\boldsymbol{\beta}})\right) + \frac{r(p+1)}{r-(p+2)},$$

where r refers to the number of uncensored observations. Secondly, we have another criterion, which is similar to AIC with exception in the penalty term. It is known as Bayesian Information Criterion (BIC) (Schwarz)[12] and gives higher penalty compared to AIC. It has the following form

$$BIC = -2log\,(L(\hat{\boldsymbol{\beta}})) + p \times log(n).$$

The penalty of BIC gives greater penalty for large number of covariates. Moreover, Volinsky and Raftery[11] extended BIC to the Cox model. They suggested using number of uncensored events instead of the number of observations. Then the corrected BIC can be written as

$$BIC_c = -2log(L(\hat{\boldsymbol{\beta}})) + p \times log(r),$$

where r refers to the number of uncensored observations.

Another popular choice for model selection is the Stepwise selection that performs variable selection. It is a hybrid technique combining the forward and backward stepwise selection strategy. Every variable is added to the model sequentially and after adding each variable, the method may discard any variable that no longer provides an improvement in the model fit. Models can be evaluated by AIC, BIC, Mallows' C_p or Adjusted R^2. Finally, there is another criterion, which also performs variable selection named Least Absolute Shrinkage and Selection Operator (LASSO). It is more sophisticated method than Stepwise selection in performing model selection. The original LASSO (Tibshirani)[9] minimizes the residual sum of squares subject to a certain constraint. For a linear regression model, the LASSO estimate $\hat{\boldsymbol{\beta}}_\lambda$ minimizes the quantity

$$\hat{\boldsymbol{\beta}}_\lambda = \min_{\beta} \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda\textstyle\sum_{j=1}^{p}|\beta_j|, \quad (2)$$

where λ is the tuning parameter and β_j is the regression coefficient in the model. LASSO was extended by Tibshirani[10] to the Cox model by using the constraint of equation (2) to the likelihood function. Then the expression becomes $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$, where $l(\boldsymbol{\beta})$ is the logarithm of likelihood for Cox model. The LASSO method requires a technique for choosing an appropriate value of λ. *Cross validation* provides simple solution for this problem. We consider a grid of λ values and choose the λ for which the CV error is the smallest.

Simulation Study

In this research, we execute a simulation study to understand how well different model selection methods perform to identify the true model. We generate the survival times, T, from the Cox model by

$$T = H_0^{-1}(-\log(U)\exp(-\boldsymbol{\beta}'x)). \quad (3)$$

To simulate survival times from equation (3), we need to specify the inverse cumulative hazard function, $H_0^{-1}(.)$. We can assume the distribution of survival times as exponential, Weibull and Gompertz distribution for Cox model (Bender et al).[2] We generate survival times from Weibull distribution in this study. The inverse of the cumulative hazard function is given by

$$H_0(t) = (\lambda^{-1}t)^{1/v}, \quad (4)$$

where λ is the known scale parameter and the v is the known shape parameter. By inserting equation (4) into equation (3), we can get the survival times of a Cox PH model with baseline hazard of a Weibull distribution, which can be expressed as

$$T = [\lambda^{-1} - \log(U) \times \exp(-\boldsymbol{\beta}'x)]^{1/v}.$$

In order to generate right censored observations, we generate a random variable c from *binomial distribution*, i.e. **c** ~ binom (1, **a**). By varying the parameter **a** in the *binomial distribution*, we can vary the censoring proportion. As for example, for 30% censoring proportion, we generate c ~ binom (1, 0.7).

III. Results and Discussion

General Settings

We consider different sample sizes (n = 30, 50, 100) with no censoring, censoring proportions (40%, 60%, 70%) and correlation coefficients (ρ = 0.2, 0.6, 0.8) among the covariates. Each dataset consists of four covariates (x_1, x_2, x_3, x_4) and the corresponding regression coefficients $\boldsymbol{\beta}'$ = (0.3,0.5,0,0). Our candidate models consist of the sequential columns of $\boldsymbol{X}$ (Hurvich and Tsai)[7]; i.e., consist of columns 1, ..., m. In addition, the true model (m_0) consists of the first 2 columns of X. The covariates are generated from *multivariate normal distribution*. The correlation between x_i and x_j, $i \neq j$ is ρ. Each of the dataset will be replicated 1000 times. We present the following tables containing the results for different criteria from the simulation study.

From Table 1, we can say that at small sample size and low correlation, stepwise method identifies a true model more often than other criteria do. However, for highly correlated dataset, LASSO shows good result compared to stepwise method (Table 3). We observe from Table 1, Table 2 and Table 3 that as the sample size increases, the rate of selecting the true model increases for both stepwise and LASSO.

AICc gives better performance than AIC in particularly small sample. However, the success rate really gets better for AIC than AICc for larger samples. Both of them are highly affected when there is censored observations in the

data. AIC and AICc both become unstable in a highly correlated data. There is no difference in the performance of BIC and BICc in absence of censoring. But classical BIC is greatly affected when censoring is considered in the dataset. BICc always shows better result than BIC in every situation. For complete and large data with weak correlation, BIC and BICc show highest success rate, 65% approximately. However, it seems BIC and BICc are greatly affected by both censoring and high correlation. The performances get worse when the covariates are highly correlated. The success rate for BIC is very low in some cases. We have not shown all the results here. If anyone feels interested to see detailed output of the simulation study may contact us.

Table 1. Results for model selection for Cox model in 1000 replications using different model selection criteria (ρ = 0.2)

Sample Size	Method	Success Rate	Average Model Size	Selected Model Order, m					
				0	1	2	3	4	m_0=2
***n*=50 censoring proportion 0%**	Stepwise	39%	1.90	0	257	502	220	21	390
	LASSO	29.7%	2.42	26	135	370	332	137	297
	AIC	39%	1.90	0	257	502	220	21	390
	AICc	35.7%	2.27	0	116	461	353	70	357
	BIC	34.3%	1.53	0	545	407	46	2	343
	BICc	34.3%	1.53	0	545	407	46	2	343
***n*=50 censoring proportion 40%**	Stepwise	26.5%	1.72	0	401	433	151	15	265
	LASSO	22.9%	2.006	104	216	338	254	88	229
	AIC	26.5%	1.72	0	401	433	151	15	265
	AICc	26.8%	1.97	0	209	447	288	56	268
	BIC	15.6%	1.29	0	734	236	29	1	156
	BICc	16.7%	1.27	0	746	237	16	1	167
***n*=50 censoring proportion 60%**	Stepwise	17.9%	1.56	0	523	357	106	14	179
	LASSO	20.1%	1.58	237	246	277	183	57	201
	AIC	17.9%	1.56	0	523	357	106	14	179
	AICc	20.5%	1.83	0	313	415	229	43	205
	BIC	0.82%	1.19	0	826	156	17	1	82
	BICc	0.94%	1.17	0	835	155	9	1	94
***n*=50 censoring proportion 70%**	Stepwise	13.5%	1.51	0	600	304	82	14	135
	LASSO	8.1%	1.37	307	258	242	141	52	81
	AIC	13.5%	1.51	0	600	304	82	14	135
	AICc	16.3%	1.91	0	368	389	208	35	163
	BIC	0.54%	1.16	0	884	104	11	1	54
	BICc	0.59%	1.16	0	857	125	14	0	59

Table 2. Results for model selection with Cox model for 1000 replications using different model selection criteria (ρ = 0.2)

Sample Size	Method	Success Rate	Average Model Size	Selected Model Order, m					
				0	1	2	3	4	m_0=2
***n*=100 censoring proportion 0%**	Stepwise	61.4%	2.24	0	75	647	245	33	614
	LASSO	42.3%	2.69	0	36	412	376	176	405
	AIC	61.4%	2.24	0	75	647	245	33	614
	AICc	43.1%	2.62	0	19	450	427	104	431
	BIC	64.8%	1.79	0	268	677	53	2	648
	BICc	64.8%	1.79	0	268	677	53	2	648
***n*=100 censoring proportion 40%**	Stepwise	52.3%	2.07	0	168	615	193	24	523
	LASSO	34.7%	2.49	11	127	374	341	147	347
	AIC	52.3%	2.07	0	168	615	193	24	523
	AICc	41.1%	2.45	0	69	491	366	74	411
	BIC	41.6%	1.54	0	496	465	39	0	416
	BICc	43.5%	1.56	0	484	477	38	1	435
***n*=100 censoring proportion 60%**	Stepwise	40%	1.90	0	304	505	177	14	400
	LASSO	29.3%	2.25	49	191	344	289	127	293
	AIC	40%	1.90	0	304	505	177	14	400
	AICc	34.2%	2.32	0	146	456	329	69	342
	BIC	24.3%	1.33	0	691	284	25	0	243
	BICc	22.5%	1.30	0	714	275	11	0	225
***n*=100 censoring proportion 70%**	Stepwise	28.9%	1.75	0	415	437	132	16	289
	LASSO	22.5%	1.96	109	233	333	241	84	225
	AIC	28.9%	1.75	0	415	437	132	16	289
	AICc	27.3%	2.20	0	204	450	290	56	273
	BIC	13%	1.20	0	813	171	16	0	130
	BICc	14.2%	1.21	0	800	191	9	0	142

Table 3. Results for model selection with Cox model for 1000 replications using different model selection criteria (censoring proportion 40%)

Sample Size	Method	Success Rate	Average Model Size	Selected Model Order, m					
				0	1	2	3	4	m_0=2
***n*=100**	Stepwise	50.3%	2.11	0	165	582	227	26	503
ρ = 0.2	LASSO	31%	2.49	1	11	37	40	11	31
	AIC	50.3%	2.11	0	165	582	227	26	503
	AICc	37.4%	2.50	0	66	448	403	83	374
	BIC	40.4%	1.51	0	522	451	26	1	404
	BICc	45.1%	1.59	0	461	492	45	2	451
***n*=100**	Stepwise	41.7%	1.84	0	305	561	119	15	417
ρ = 0.6	LASSO	36.9%	2.54	3	92	397	375	133	369
	AIC	41.7%	1.84	0	305	561	119	15	417
	AICc	37.4%	2.26	0	138	525	273	64	374
	BIC	16.6%	1.25	0	761	229	10	0	166
	BICc	20.2%	1.33	0	696	281	23	0	202
***n*=100**	Stepwise	17.9%	1.53	0	558	358	78	6	179
ρ = 0.8	LASSO	33.9%	2.42	0	120	428	366	86	339
	AIC	17.9%	1.53	0	558	358	78	6	179
	AICc	22%	1.23	0	260	470	228	42	220
	BIC	2.2%	1.07	0	939	51	10	0	22
	BICc	3.1%	1.10	0	909	81	10	0	31

Real world example: BDHS 2014 data

We have used Bangladesh Demographic and Health Survey (BDHS) 2014 data to evaluate the performance of different criteria. We have considered 5099 observations in this study. There are 383 events and the censoring proportion is almost 92%. As we intend to observe infant mortality, our response variable is the death of a child before its first birthday. Therefore, the event of interest occurs if the death is before 12 months. The censoring indicator is: 1 = death before 12 months, 0 = otherwise. The potential covariates in this study include the following variables: Mother's age (x_1), Region (x_2): Barisal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur and Sylhet, Type of residence (x_3): Urban and Rural, Mother's education (x_4): No education, Primary education, Secondary education and Higher education, Sex of the child (x_5), Mother's occupation (x_6), Wealth Index (x_7).

Table 4 shows the results after fitting the Cox model including all the covariates considered in this study.

The estimated coefficients, their standard errors and $p-$ *values* are presented in the table. Table 5 represents the result for variable selection by using stepwise selection and LASSO. Table 6 shows the results for model selection by using AIC, AICc, BIC and BICc under Cox model. The selected models, estimated coefficients of the parameters are presented here.

Table 4 shows, variable x_1 (mother's age) has the significant impact on the death of a child. Then the variable x_2 (sex of child) has also influence on child's death.

Table 4. Results for full model for BDHS 2014 data under Cox model

Covariates	Value	Std. Error	z	*p*-value
x_1	-0.071	0.007	-9.83	0.000
x_2	-0.049	0.032	-1.58	0.115
x_3	0.136	0.115	1.18	0.237
x_4	-0.057	0.061	-0.94	0.347
x_5	0.207	0.102	2.03	0.043
x_6	0.005	0.004	1.25	0.210
x_7	0.044	0.041	1.08	0.282

Table 5. Results for variable selection using stepwise and LASSO for BDHS 2014 data under Cox model

Covariates	Stepwise	LASSO
x_1	-0.069	0.022
x_2	0.000	0.000
x_3	0.000	0.000
x_4	0.000	0.000
x_5	0.207	0.000
x_6	0.000	0.000
x_7	0.000	0.000

Table 5 shows, stepwise method selects x_1 as well as x_5 as the most significant predictors. However, LASSO selects only x_1 as the significant predictor. From Table 6, we observe that AIC selects the model with the covariates (x_1, x_5) as the best one. AICc selects the model with covariates (x_1, x_2, x_3, x_5, x_6, x_7) as the best one. BIC and BICc both select the model with covariate x_1 as the best one. The model suggested by AIC shows that the estimated parameters are statistically significant. The $p-$ *value* of the model selected by BIC and BICc is also statistically significant. If we see the model selected by AICc, then it can be noticed that all other covariates except x_1 and x_5 are insignificant here.

Table 6. Results for model selection using AIC, AICc, BIC and BICc for BDHS 2014 data

Criterion	Covariates	Value	Std. Error	z	*p*-value
AIC	x_1	-0.069	0.0069	-10.01	0.00
	x_5	0.207	0.102	2.03	0.043
BIC, BICc	x_1	-0.069	0.0069	-10	0.00
	x_1	-0.069	0.0069	-9.99	0.00
	x_2	-0.0356	0.029	-1.21	0.225
AICc	x_3	0.129	0.114	1.13	0.259
	x_5	0.209	0.102	2.04	0.041
	x_6	0.004	0.004	1.08	0.282
	x_7	0.045	0.041	1.09	0.274

IV. Conclusion

In this research, we have evaluated the performance of different model selection criteria with Cox model. AIC gives more stable result than AICc when the sample size is large. BICc is better than BIC to identify the true model in every setting. However, it is difficult to choose one best method since all the methods have their advantages and disadvantages. We observe how their performances become volatile under different extreme situations (e.g. small sample size, heavy censoring and high correlation among the covariates. Therefore, the decision to choose the appropriate method in selecting a true model has to be made with great caution. This research can further be extended by considering interactions among the covariates along with other survival regression models.

References

1. Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC-**19**, 716-723.
2. Bender, R., T. Augustin and M. Bletter, 2005. Generating Survival Times to Simulate Cox Proportional Hazards Models. Statistics in Medicine, **24**, 1713-1723.
3. Collett, D., 2003. Modelling Survival Data in Medical Research. Chapman and Hall; New York.
4. Cox, D.R., 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, series B;* **34**, 187-220.
5. Efroymson, M.A., 1960. Multiple regression analysis. *Mathematical methods for digital computers,* **1**, 191203.
6. Hurvich, C.M., J.S. Simonoff and C.L. Tsai, 1998. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *Journal of the Royal Statistical Society, Series B,* **60**, 271-293.
7. Hurvich, C.M. and C.L. Tsai, 1989. Regression and time series model selection in small samples. *Biometrika,* **76**, 297-307.
8. Therneau, T.M. and P.M. Grambsch, 1998. Penalized Cox models and frailty. *Technical report, Division of Biostatistics. Mayo Clinic; Rochester,* **12**, 156-175
9. Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B,* **58**, 267-88.
10. Tibshirani, R., 1997. The lasso method for variable selection in the Cox model. *Statistics in Medicine,* **16**, 385-95.
11. Volinsky, C.T. and A.E. Raftery, 2000. Bayesian information criterion for censored survival models. *Biometrics,* **56**, 256-62.
12. Schwarz, G. 1978. Estimating the dimension of a model. The annals of statistics, **66**, 461-464.