ISSN: 0974-5823 Vol. 7 No. 4 April, 2022

International Journal of Mechanical Engineering

ROBUST MODEL SELECTION WITHIN THE COX MODEL USING BLADER CANCER DATA

[1]Sasikala. S and [2]Karunanidhi. D

[1]Assistant Professor, Department of Statistics, Thanthai Periyar Govt. Arts and Science college (Autonomous) (Affiliated to Bharathidasan University) Tiruchirappalli-23.

[2]Research Scholar, Department of Statistics, Thanthai Periyar Govt. Arts and Science college (Autonomous) (Affiliated to Bharathidasan University) Tiruchirappalli-23.

Abstract

Bladder cancer is one of the supreme communal cancers universally and available analytical models and techniques are derisory to know the performance of the diseases. In Statistical modeling, Model Selection method is one of the important statistical processes. This study focusing on Cox model which is characterized into two criterion model choice such as Schwartz Akaike criterion and Robust Akaike Criterion. The criterions are based on a smooth modification of the partial likelihood function. Asymptotic results are presented in this study and also Monte Carlo study has been obtained to shows the finite sample behavior of the procedure under divergences from the Cox model.

Keywords: Model Selection, Akaike's Criterion, Schwarz Information Criterion, Cox Model, Monte Carlo Methods, Partial Likelihood function.

INTRODUCTION:

Bladder cancer is one of the most recurrently happening human cancers, emergent through two tracks such as papillary and non -papillary that analogous to clinically different forms of the disease. Most bladder cancers are chemically tempted, with tobacco smoking being leading the risk factor. The occurrences of the bladder cancer more in the European countries and specifically in Italy and Spain identified a greater number of bladder cancer cases. Bladder cancer is the cancer which grows and develops in the bladder lining. In some of the cases, the tumor spreads into the bladder muscle. Bladder cancer is the fourth most common cancer in men and the eighth most common in women. The prevalence rate is about four times greater among men than women.

Bladder cancer performances is based on the stages and risk of the cancer recur further. Bladder cancer stagging is based upon the how far it is infiltrated into the tissues of the bladder and the cancer involves lymph nodes near the bladder, and whether the cancer has spreads beyond the bladder to the other organs. The verdict stage is the most important predictive factor for invasive bladder cancer, while grade is the most important predictive factor for non-invasive bladder cancer. Based on the stage, most of the studies uses the abridged cataloguing based on localized, regional, and distant. When reconnoitering survival outcomes for patients with bladder cancer, most of the studies rely on conservative statistical methods such as proportional hazards models. In this paper, gene expression and clinical data related to bladder cancer were obtained from TCGA and GEO databases and Cox model used for analyses the data. These analytical models such as Schwarz, Akaike and Bayesian information criterion based on the partial likelihood function are established to study about the bladder cancer. The robust models are used to predict the effectiveness of the patients who all are taking individualized treatments.

Different number of procedures were projected for the estimation in the Cox model (Cox, 1972) through the functional approach. The Cox regression model (Cox, 1972) appropriate to study how the unemployment time influences the socio-economic factors and also described the distribution of the factors. The model choice in regression models explores the explanatory variables using the number of process and attains the results in the final step of the processes. The Schwarz information criterion (Schwarz, 1978) is the most superior tools for model selection of data analysis. It is based on minimization of $-\log L(j, \hat{\theta}^j) + 0.5 d(j) \log n,$ where $L(\cdot)$ is the likelihood function, j indexes sub-models, $\hat{\theta}^j$ is the maximum likelihood estimator of θ in the j^{th} model, $d(j)$ is the dimension of the model while n is the number of independent observations. The SIC more appropriate when the inference based on the objective functions in robust estimation [10]. In [4] the modification of partial likelihood function has been introduced in order to achieve robustness of estimation procedure. These estimation procedures are more useful to describes about the Cox model.

Cox models are developed conventional statistical approach which resulted with efficient parametric (or semi-parametric) procedures with its robust counterparts. There are models, like the normal one, where those robust procedures can be routinely provided. Due to complex structure of the partial likelihood function of Cox models, the estimator which is evaluated using the model is efficient. First fix basic notations to explain the merits of robust estimation. Denote by F_n the empirical distribution function of the random sample (T_1, C_1, Z_1). . . (T_n, C_n, Z_n), where T, C and Z are, respectively, time, censoring and covariate variables. Censoring and time are independent, given the value of the covariates–explanatory variables. Under the Cox model observed are only $(\tilde{T}_i, \delta_i, Z_i),$ where $\tilde{T}_i$ is the minimum of the time variable *T* and censoring *C*, while δ_i indicates $\tilde{T}_i = T_i,$, for *i* = 1, . . . , *n*. Using Cox model, Robust estimators of the regression parameter are considered in [4] to solve the following equation:

$$\int A(w,y)\left[y - \frac{\int A(w,y) z I_{\{\min(a,t)\geq w\}} \exp(\beta' z) d\ F_n(t,a,z)}{\int A(w,y) I_{\{\min(a,t)\geq w\}} \exp(\beta' z) d\ F_n(t,a,z)}\right] I_{\{w\leq c\}} d\, F_n(w,c,y) = 0 \qquad ...(1)$$

equivalent to the partial likelihood score function equation

$$\int\left[Z_i - \frac{\sum_{Min(T_j,C_j)\geq T_i} Z_j \exp(\beta' Z_j)}{\sum_{Min(T_j,C_j)\geq T_i} \exp(\beta' Z_j)}\right] I_{T_i \leq C_i} = 0,$$

When the function *A* weight function is equal to 1.

While the first-order equation (1) involves explicitly the unobservable empirical Cumulative Distribution Function (CDF) F_n, specific form of integrated functions ensures that the left-hand side is computable under the observed sample. Solving the equation with respect to β, if the solution exists, is equivalent to maximization of the following modified partial likelihood:

$$\prod_{i=1}^{n}\left\{\frac{\exp(\beta' Z_i)}{(1/n)\sum_{\min(T_j,C_j)\geq T_i}\left[A(T_i, Z_J)\exp(\beta' Z_j)\right]}\right\}^{\delta_i A(T_i,Z_i)} \qquad ...(2)$$

The logarithm of the modified likelihood, which is concave with respect to β, can be written

conveniently as

$$L_{F_n}, A(\beta) = \int I_{w\leq c}\ A(w,y)\left[\beta' y - \log\int I_{\min(t,a)\geq w}\ A(w,z)\exp(\beta' z) d\ F_n(t,a,z)\right] d\ F_n(w,c,y) \qquad ...(3)$$

To assure the robustness and proper asymptotic behaviour of the new estimator of weights and the estimators should be smooth and down-weight outlying observations. This methodology has been pointed out in [3] that observations with excessive values $T_i e^{\beta' Z_i}$, where β is the true parameter value, are most influential in the inference process. More formally, solutions to (1), when weights are 1, may converge to arbitrary values with respect to sequences of distributions approaching, in sup norm for cdf, a Cox model distribution. It was also argued there, that an important source of instability is in violation of the dependence structure between *T* and *Z* thinkable as 'erroneous' data in practical situations. Therefore, a natural choice for the weights *A* was in down-weighting the largest values of $\tilde{T} e^{\beta' Z}$. [4] has shown that it is possible to choose families of *A* functions so that the estimator functional becomes uniformly Frechet differentiable at the Cox model distributions. The fact guarantees robustness of the implied estimators in the usual sense, the uniform normal approximation of the estimator's distribution with respect to small (infinitesimal) non-parametric neighborhoods of the model law, as well the reliability in estimator's variance assessment.

MONTE CARLO STUDY

Computations of robust estimator, it is an essential one for the robust model selection method, are based on the following steps (compare [4]):

1. Find the partial likelihood estimator $\hat{\beta}$
2. Evaluate 0.9 (or 0.95) empirical quantile for the sample $\tilde{T}_1 \exp(\hat{\beta}' Z_1), ..., \tilde{T}_n \exp(\hat{\beta}' Z_n)$ and denote it by *M*,
3. Maximize (2) for the weights

$$A(t,z) = M - \min\{M, t\ \exp(\hat{\beta}' Z)\},$$

4. Estimate the cumulated hazard using [8] robustified version of Breslow's estimator

$$\hat{\Lambda}_A(t) = \sum_{T_i \leq t} \frac{A(T_i, Z_i)\,\delta_i}{A(T_i, Z_i)\exp(\hat{\beta}' Z_j)} \qquad ...(5)$$

5. Compute, as in point 2, the empirical quantile *M* for the sample

$$\hat{\Lambda}(T_1)\exp(\hat{\beta}' Z_1), ..., \hat{\Lambda}(T_n)\exp(\hat{\beta}' Z_n)$$

and go to 3 till stability is achieved.

Robust estimation is the method which is follows the strong differentiability property of the estimation or functional. Specific choice of the family of weight functions is based on the fact (see [3], [4]) that stability in the partial likelihood estimation comes with observations which are $\widetilde{T}_i e^{\beta'_0 Z_i}$ excessive. Weights given above reduce this influence. Notice that the random variable $\Lambda(T) e^{\beta'_0 Z}$ is standard exponential under the Cox model, while it is reasonable to expect $T e^{\beta'_0 Z}$ be approximately exponential when (t) is 'nearly' linear in T. Adaptively, of the method based on weights given in point 3 of the algorithm was shown in [5]. Extension to weights based on $\hat{\Lambda}(T) e^{\hat{\beta}'_0 Z}$ is natural (not yet formally supported) and shown, by the Monte Carlo experiments, to be a little more efficient.

The weights alteration is regularly attained after 3 to 4 iterations. The weight functions A (t, z) given above does not satisfy all required formal regularity conditions. The smoothed version multiplied by selective function A_0 has given virtually unchanged results for mild in covariates contaminations. To meet all the formal assumptions, one has to multiply the above smoothed weights by a smooth compact support function of z excluding observations with improbable covariates the step usually accomplished by proper scanning of data. Alternatives for the weights were also considered, for instance.

$$A(t,z) = \exp\{-\Lambda(t) e^{\beta' z} / (\alpha M)\} \ldots (6)$$

Where, α is a scaling factor. They showed comparable behaviour.J

OBJECTIVES OF THE STUDY:

- To find the best diagnostic the bladder cancer using Schewarz, Akaike's and Bayesian Criterion.
- To evaluate the concert of the models using robust estimators.
- To estimate the robust estimators using Monte Carlo Study.

DATA SET AND METHODOLOGY

In this segment, the real-life data has been analyzed which is taken from the National Cancer Institute Surveillance Epidemiology and End Results (SEER) database covers data. The data consists of 18 population-based registries and represents approximately 28% of the population [12]. The SEER case citation session was used to identify patients from the SEER 18 (November 2018 submission) database. In this study, the explanatory variables are age, sex and Stage I, Stage II, Stage III, Stage IV and No stage on each specific type of cancer affects a particular type of cell present in the blood. A repetitive blood test detecting the occurrence of these type of cancers as early as possible. The statistical analysis might carry out for the given data. The model selection and estimation are presented in the consequent tables are supposed to determines the excessive consideration of the standard approach on model selection.

The first column of each table contains variable names. The second one gives MLE and columns under the heading 'SE' give estimated standard errors.

Table.1. Comparison of model selections : Schwarz, Akaike information criterion and Bayesian information criterion based on the partial likelihood for the bladder cancer for the uncontaminated data (robust criterion corresponding to 0.9 Quantile).

Regressors	**MLE**	**S.E.**	**AIC**	**S.E**	**BIC**	**S.E**	**SIC**
Sex	2.965	0.328					
Age	1.156	0.421	123.456	0.032	125.3291	0.041	125.3291
Stage I	1.762	0.314	121.765	0.292	124.651	0.256	138.808
Stage II	1.888	0.287	119.245	0.354	123.045	0.312	124.584
Stage III	1.984	0.396	120.621	0.354	124.003	0.308	134.467
Stage IV	1.879	0.386	121.002	0.975	125.513	0.751	130.943
No Stage	1.751	0.264	118.64	0.217	123.107	0.213	139.852

Table.2. Comparison of model selection: Schwarz, Akaike information criterion and Bayesian information criterion based on the partial likelihood for the bladder cancer for the contaminated data (robust criterion corresponding to 0.9 Quantile).

Regressors	MLE	S.E.	AIC	S.E	BIC	S.E	SIC
Sex	2.812	0.488					
Age	1.261	0.127	123.862	0.184	126.915	0.141	16.917
Stage I	1.818	0.347	122.007	0.327	125.581	0.296	139.014
Stage II	1.907	0.297	119.455	0.397	124.475	0.362	125.842
Stage III	1.991	0.412	120.981	0.384	124.984	0.387	134.977
Stage IV	1.979	0.401	121.902	0.981	126.103	0.793	131.023
No Stage	1.723	0.316	119.514	0.305	124.707	0.259	138.522

In Table 1, Schwarz, Akaike and Bayesian criterion are compared with respective Standard Error. The data analyzed and compared partial likelihood estimation with different model selections. In Table 2, shows the effect of contamination data which is based on the robust measures. After many repetitions of robust model selection, the stability has been attained and also the model selection based on Schwarz method of an uncontaminated to contaminated data. The robust Akaike method remain unchanged virtually for both the data sets.

CONCLUSION:

The survival rate of the patients who are having bladder cancer may vary momentously of various stages among both non-invasive and invasive cases. The non-invasive cancers are more highly affectable one among human when compared with invasive cases. Moreover, those factors are represented independently. In survival data, sex has no association with any other factors and Age and stage are significant. This can be used to predict patients' survival outcomes. In this study, partial likelihood method and robust method are used for model selection. The robust model selection given the highest stability for the given data set using the Schwarz criterion. The Robust methods are the best methods to evaluate the successful rates of bladder cancer treatments with the robust estimators. The accuracy of predictions values based on robust models may verify in further studies and also studies the clinical applications of the individualized treatment of bladder cancer.

References

1. Akaike, H. (1969). Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics 21, 243–47.
2. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory. Budapest: Academiai Kiado pp. 267–81.
3. Bednarski, T. (1989). On sensitivity of Cox's estimator. Statistics and Decisions 7, 215–28.
4. Bednarski, T. (1993). Robust estimation in Cox's regression model. Scandinavian Journal of Statistics 20, 213–25.
5. Bednarski, T. (1999). Adaptive robust estimation in the Cox regression model. Biocybernetics and Biomedical Engineering 19(4), 5–15.
6. Cantoni, E. and Ronchetti, E. (2001a). Robust inference for generalized linear models. J. Amer. Statist. Assoc. 96 1022-1030.
7. Cox, D. R. (1972). Regression models and life tables. Journal of the Royal Statistical Society, Series B 34, 187–20.
8. Grzegorek, K. (1993). On robust estimation of baseline hazard under the Cox model and via Fr´echet differentiability. Institute of Mathematics, Polish Academy of Sciences. Preprint No 518.
9. Kaufman DS, Shipley WU, Feldman AS (2009). Bladder cancer. Lancet. 374(9685), 239–49.
10. Machado, J. A. F. (1993). Robust model selection and M-estimation. Econometric Theory 9, 478–93.
11. Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6(2), 461–64.
12. Surveillance Epidemiology and End Results program (www.seer.cancer.gov). Database: Incidence: SEER 18, November 2018 submission.