International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 26th Dec 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue12

10.48047/IJIEMR/V11/ISSUE 12/230

TITLE: A STUDY OF VARIABLE SELECTION FOR FRAILTY AND MARGINAL MODELS IN SURVIVAL ANALYSIS

Volume 11, ISSUE 12, Pages: 1758-1768

Paper Authors **S NIRMALA, DR. RAM BALI SINGH**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

A STUDY OF VARIABLE SELECTION FOR FRAILTY AND MARGINAL MODELS IN SURVIVAL ANALYSIS

S NIRMALA, DR. RAM BALI SINGH

DESIGNATION- RESEARCH SCHOLAR MONAD UNIVERSITY HAPUR U.P
DESIGNATION- (PROFESSOR) MONAD UNIVERSITY HAPUR U.P

ABSTRACT

The research also investigates the effect of several tuning factors on the variable selection procedure, such as the regularization strength in LASSO and the quantile level in LAD regression. Furthermore, the approaches' susceptibility to multicollinearity and missing data is explored. In this study, we show that LASSO and Least Absolute Deviation (LAD) regression are equally useful for variable selection in survival analysis. By forcing certain coefficients to zero, LASSO creates sparse models, whereas LAD regression offers robustness against outliers and is hence well-suited to datasets with potentially noisy observations. It is important to consider the peculiarities of the dataset and the researcher's interest in interpretability when deciding which approach to use. In sum, this research adds to the body of knowledge on survival analysis by shedding light on how LASSO and LAD regression might be used for variable selection. In areas where time-to-event data analysis is crucial, researchers and practitioners may use these results to improve the accuracy and interpretability of their survival models.

KEYWORDS: Variable Selection, Frailty and Marginal Models, Survival Analysis, LAD regression, LASSO creates sparse models.

INTRODUCTION

Survival analysis relies heavily on careful variable selection. In actual clinical practice, a large number of confounders may be included. Data scientists often include several predictors in the early stages of model development. A more parsimonious model is preferable in all cases since it improves model prediction and interpretation. Consequently, in the presence of a high number of predictors, identifying important variables plays key roles in model development and is quite difficult. In high-dimensional statistical modeling, variable selection is crucial. For linear regression models, several authors have presented different variable selection criteria and approaches. Due to the complex nature of the data, the study of

variable selection in survival analysis has attracted a lot of interest in recent years.

In statistics, there is a subclass of survival models known as Cox's Proportional Hazards models. Time to an event is the dependent variable in survival models, which also take into account other possible factors. Each covariate's influence on the hazard rate is multiplicative in a proportional hazards model, rather than additive. The Cox proportional hazards model takes for granted the independence of subject survival times. However, when the data are correlated, it is possible that this assumption is not met. In this case, the independence assumption among individuals is broken; hence the widely used Cox model cannot be used. The frailty model and the marginal model are two extensions of the Cox regression model for the study of multivariate failure time data. It is planned to go through how to choose variables for marginal and frailty models in survival analysis in this section. Real-world examples are used to back up the numerical findings.

HARD Threshold Penalty

Fan noticed the punishment function for the penalized least-squares estimator

$p(|\theta|) = |\theta|I(|\theta| \leq \lambda) + \lambda/2I(|\theta| > \lambda)$ leads to the hard-thresholding rule

$\theta^{\wedge} = zI(|z| > \lambda)$

This penalty function does not excessively punish the case when $|\theta|$ is really big. Fan recommended the following severe punishment for crossing a threshold:

$p\lambda(|\theta|) = \lambda 2 — (|\theta| — \lambda)2I(|\theta| < \lambda)$ With the clipped L1-penalty function $p\lambda(|\theta|) = \lambda \min(|\theta|, \lambda)$

The solution is a mixture of soft and hard thresholding rule $\theta^{\wedge} = sgn(z)(|z| — \lambda)+I(|z| \leq 1.5\lambda) + zI(|z| > 1.5\lambda)$

Least Absolute Shrinkage and Selection Operator (LASSO)

Assume we have (xi,yi) data, where xi=(xi1,..., xip)T are predictor variables and yi are the answers, for some integers i=1, 2,..., N. We assume either independence of observations or conditional independence of yis given xijs, as in the standard regression setup. For the sake of argument, let's say the xij are all the same: $\Sigma_i x_{ij}/N = 0, \Sigma_i x_{ij}^2/N = 1.$

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Letting $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$, the lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined by

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2 \right\} \text{ subject to } \sum_j \left|\beta_j\right| \leq t$$

Here $t \geq 0$ is a tuning parameter. For all t, the solution for α is $\alpha\hat{} = y$. For a detailed study refer to Tibshirani (1996).

Generalized Cross Validation (GCV) estimate

The generalized Cross Validation estimation of λ is the minimizer of $V(\lambda)$

$$V(\lambda) = \frac{(1/n)\left\|(I - A(\lambda))y\right\|^2}{\left[(1/n)\mathrm{tr}(I - A(\lambda))\right]^2}$$

where $A(\lambda)$ is the $n \times n$ influence matrix, which satisfies

$$\begin{pmatrix} f_{n,\lambda}(t_1) \\ \vdots \\ f_{n,\lambda}(t_n) \end{pmatrix} = A(\lambda)y, y = (y_1, \ldots, y_n)$$

The GCV estimates the λ which minimizes the predictive mean square error $R(\lambda)$ defined by

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left(f(t_i) - f_{n,\lambda}(t_i)\right)^2$$

$f_{n,} \lambda (t)$, $t \in [0,1]$ is also a Bayes estimate of f(t) if f has a certain partly incorrect zero-mean Gaussan prior.

Cox's Proportional Hazard Model

Let $t_1^0 < \cdots < t_N^0$ Rank the reported occurrences of failure. Let x(0), x(1),..., x(N) be the variables associated with the N successes, and let (j) represent the label for the item that fails at t0. Let's call the danger you took before time t0 "Rj.:

$$R_j = \left\{i: Z_i \geq t_j^0\right\}$$

The formula for the Cox proportional hazards model is,

$$h(t|\mathbf{x}) = h_0(t)\exp(x^T\beta) \quad \ldots.(3)$$

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

parameters and h0(t), which represent the baseline hazard function. The probability described in eq. (2) gets more likely the earlier it is discussed.

$$L = \prod_{i=1}^{N} h_0(Z_{(i)}) \exp(x_{(i)}^T \beta) \prod_{i=1}^{n} \exp\{-H_0(Z_i) \exp(x_i^T \beta)\}$$

where H0(.) is the cumulative hazard function at the first state. For a parametric baseline hazard function, h0(,.), the associated penalized log-likelihood function is

$$\sum_{i=1}^{N} [\log\{h_0(\theta, Z_{(i)})\} + x_{(i)}^T \beta] - \sum_{i=1}^{N} \{H_0(\theta, Z_i) \exp(x_i^T \beta)\} - n \sum_{j=1}^{d} p\lambda(|\beta_j|) \quad \dots (4)$$

The maximum penalized likelihood estimator is found by optimizing eq. (4) with respect to (θ, β).

Pseudo Likelihood

The joint probability distribution of a set of random variables is approximated by the concept of pseudo likelihood. An approximation to the likelihood function of a collection of observed data may be obtained, which can be used to either provide explicit estimates of model parameters or to simplify the estimation issue computationally.

Given a set of dependencies E between random variables, where Xi, Xj E denotes Xi is conditionally independent on Xj given Xi's neighbors, the pseudo likelihood of X = (x1, x2,..., xn) is the product of the probabilities of each of the random variables in X.

$$\Pr(X - x) - \prod_i Pr(X_i - x_i | X_j - x_j \; for \; all \; j \; for \; which \; \{X_i, X_j\} \epsilon E)$$

The vector of variables denoted by X, and the vector of values denoted by x. Each element of the vector X has a corresponding value in the vector x, as shown by the statement X = x. The chance that the vector of variables X is equal to the vector x is represented by the equation Pr(X = x). Given that state variables may be used to characterize many circumstances, the probability of a given state, Pr(X = x), can be expressed as a percentage of all possible states, given the values of the state variables. The above equation may be used to calculate a related metric called the Pseudo-log-likelihood. Thus

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

$$log\ Pr(X = x) = \sum_i log\ Pr(X_i - x_i | X_j - x_j\ for\ all\ \{X_i, X_j\} \epsilon E)$$

The pseudo-likelihood of an assignment to Xi may often be computed more efficiently than the likelihood, especially when the latter may require marginalization over a large number of variables, making it useful as an approximation for inference about a Markov or Bayesian network.

Error in predictions and model fit

If (x) is a prediction technique developed using the current data and x is a random covariate, then the prediction error is defined as

$$PE(\hat{\mu}) = E\{Y - \hat{\mu}(x)\}^2$$

where the forecast is made simply in light of the latest data point (X,Y). There is a way to break down the prediction error:

$$PE(\hat{\mu}) = EVar(Y|x) + E\{(Y|x) - \hat{\mu}(x)\}^2$$

Stochastic mistakes are the root cause of the first part. The second part stems from a fundamental model not being properly implemented. This factor is indicated by the symbol () and is known as the model error. Cox's proportional hazards regression model (4.3)

$$\mu(x) = E(T|x) = \int_0^\infty h_0(t) exp(x^T \beta) exp\left\{ - \int_0^t h_0(t) exp(x^T \beta) du \right\} dt$$

It shall be assumed that h0 (t) 1 in the simulation examples that follow. So, according to a rough algebraic estimate,

$$\mu(x) = exp(x^T \beta)$$

For the Cox frailty model with h0(t) ≡ 1,

$$\mu(x) = exp(x^T \beta)\, E(u^{-1})$$

When comparing the efficacy of two methods, we may eliminate the fragility-causing

component E(u-1) by calculating their Relative Model Errors (RME), which is just the ratio of their respective model errors. Therefore, we shall define the model error as

$$E\{exp(-x^T\hat{\beta}) - exp(-x^T\beta_0)\}^2$$

in accordance with either the Cox or the frailty model.

Instance of actual data

The Frailty Model (i)

The Government Hospital in Karaikal town, Government of Puducherry, is used to test the suggested frailty model, which is based on the work of Morris et al. (1994). Here's a detailed breakdown of what's included in this dataset:

x_1 – treatment indicator

$$x_1 = \begin{cases} 1 \text{ if treated at a nursing home} \\ 0 \text{ otherwise} \end{cases}$$

x_2 – variable mother age

$x_2 = \{k \text{ such that } k \in (25,40)\}$

x_3 – gender

$$x_3 = \begin{cases} 1 \text{ if Male baby} \\ 0 \text{ if Female baby} \end{cases}$$

x_4 – birth weight

Three binary indices of health state, x5, x6, and x7, range from greatest health to worst health. Morris et al.'s (1994) proposed model is

$$h(t|x) = h_0(t)\exp\left(\sum_{i=0}^{7} x_i\beta_i\right)$$

Where h0 (t) is the baseline hazard function utilizing gamma fragility and the Lin (1993) technique we explained before. Three parametric and one nonparametric baseline hazard models are used to fit the Cox model to this data set. Other variables are binary; therefore only x2 is standardized. The SCAD, L1, and harsh penalty of the penalized partial likelihood method are applied to this dataset. The GCV chooses the thresholding parameters of 0.02 for the SCAD, 0.01 for the LASSO, and 0.09 for the HARD. In addition, the AIC and BIC for the optimal subset variable selection are calculated. Parameter estimates have been calculated using the method and software proposed by Lin (MULCOX, 1990)*, Lin (MULCOX2,

1993**).

Marginal Model (Case II)

The National Eye Institute's Diabetic Retinopathy research (1981) aimed to determine whether or not laser photocoagulation might effectively postpone blindness in people with diabetic retinopathy.

The Vinayaka Mission's Medical College and Hospital in rural Karaikal, Puducherry, undertook a study to determine the prevalence of cataract blindness in male and female patients in 2015.

Diabetic retinopathy has been detected in this patient population. A total of 78 participants enrolled between January 2014 and December 2015.

Data were obtained from Vinayaka Mission's Medical College and Hospital using a methodology similar to that of Huster et al. and Liang et al., in which one eye of each patient was randomly chosen for photocoagulation and the other eye was examined without treatment.

Patients were monitored to detect any instances of unilateral blindness. Some degree of interdependence between the patient's two eyes is to be expected.

Think about the scenario described by the equation (4.10): Zik = (Z1ik, Z2ik, Z3ik) (i = 1,...,126; k = 1,2)

$$Z_{1ik} = \begin{cases} 1 \text{ if the kth eye of the ith patient was on treatment,} \\ 0 \text{ otherwise;} \end{cases}$$

$$Z_{1ik} = \begin{cases} 1 \text{ if the } i^{th} \text{ patient adult onset diabetes,} \\ 0 \text{ if the } i^{th} \text{ patient juvenile onset diabetes;} \end{cases}$$

and $Z_{3ik} = Z_{1ik} \times Z_{2ik}$.

COX REGRESSION MODEL VARIABLE SELECTION FOR LUNG CANCER AND LIVER DATA

An essential stage in evaluating survival data to identify the variables affecting the time to an event of interest, such as death or disease progression, is variable selection for Cox's regression model utilizing data from lung cancer and the liver. To examine the association

between variables and survival times while taking censored data into account, survival analysts often turn to Cox's proportional hazards regression. This review will go into the methods used to identify important predictors in Cox's regression model, the significance of variable selection, and the unique difficulties and concerns presented by data on lung cancer and the liver.

Cox's regression model relies heavily on careful variable selection to isolate the most important predictors of the outcome of interest and leave out any others that are unnecessary to the analysis.

Overfitting occurs when a model has too many parameters that cannot be explained by the data at hand, which might introduce bias. However, if critical variables are left out of the analysis, we may be left with inaccurate estimates and less information to work with, which will impede our ability to determine what factors influence survival rates. Improving the analysis's predictive ability and scientific worth is possible via careful variable selection, which yields models that are both parsimonious and interpretable.

Challenges and Considerations with Lung Cancer and Liver Data:

1. there is the issue of censored data, which arises in survival analysis when some people reach the conclusion of the research without having experienced the event of interest. Censored observations are more common in data involving lung cancer and liver disease since some patients may be alive or in remission when the trial is over. It is crucial to handle censored data correctly during variable selection in order to prevent inaccurate model estimates and biased findings.
2. high-dimensional data is becoming more common as a result of technological developments in data collecting. This is especially true for lung cancer and liver datasets, which often include a significant number of predictors. Problems arise in variable selection when working with high-dimensional data because of the increased computing complexity and the possibility of misleading connections.
3. multicollinearity, which occurs when predictors are strongly associated with one another, makes it more difficult to isolate the effects of individual predictors and generates unstable coefficient estimates.

Best Practices for Variable Selection in Cox's Regression Model with Lung Cancer and Liver Data:

1. First, the data must be preprocessed thoroughly to account for missing values, normalize variables, and deal with outliers that may compromise the efficiency of variable selection.

2. To check how well the model does on new data and avoid overfitting, it's a good idea to use cross-validation methods like k-fold cross-validation.

3. Regularization Strength Choosing the right regularization strength (penalty parameter) is critical in regularization methods like LASSO. To find the sweet spot between model complexity and prediction performance, cross-validation may be used.

4. when working with high-dimensional datasets, it is important to include domain expertise and previous research results to help guide the selection of suitable predictors.

Choosing which variables to include is a vital part of developing reliable and interpretable Cox's regression models for survival data including lung cancer and liver disease. Common approaches include univariate analysis, stepwise selection, regularization methods, information criteria, and RFE. Censored data, high-dimensional data, and multi collinearity are just a few of the obstacles that researchers must overcome to get somewhere useful. The variables affecting survival outcomes in studies of lung cancer and the liver may be better understood if researchers adhere to best practices for identifying important predictors, improving model performance, and gaining meaningful insights.

CONCLUSION

The numerical examples if the distributions are properly described, the marginal model is more effective than the frailty model. Results for new researchers working in the domain of variable selection in survival analysis are improved, and it is concluded that picking significant variables plays key roles in model construction and is particularly tough in the presence of a large number of predictors. When compared to stepwise selection, the LASSO

method for choosing variables in the Cox model seems to hold its own. It's more consistent than the incremental method while still producing understandable models. In order to ensure that the LASSO method's penalization scheme is uniformly applied to all regressors, it is necessary to standardize them beforehand. The LASSO selects about the right amount of zero coefficients, and it significantly beats stepwise selection. When compared to previous methodologies, the suggested methodology employed in this paper—which is based on data on lung cancer and liver cancer—performs very well. The study in this research has concentrated on fixed covariates; however time dependent covariate may be included with no more work.

REFERENCES

1. Austin, P. C. (2016). Variance Estimation when using Inverse Probability of Treatment Weighting (IPTW) with Survival Analysis. Statistics in Medicine, Vol. 35, Issue. 30, pp. 5642-5655.

2. Badrinathi, R., and Tiwari, R. C. (1992). Hierarchical Bayesian Approach to Reliability Estimation Under Competing Risk. Vol. 32, Issue 1-2, pp.249-258.

3. Baker, S.G. (1994). Regression Analysis of Grouped Survival Data with incomplete Covariates: Nonignorable Missing data and Censoring Mechanisms. Biometrics, 50, 821-826.

4. Balakrishnan, N., et al. (2006). Advances in Distribution Theory order Statistics, and Inference, Birkhauser, Berlin.

5. Bedrick, E.J. and J.R. Hill, (1996). Assessing the Fit of the Logistic Regression Model to Individual Matched Sets of Case-control data, Biometrics, 52, 1-9.

6. Bhattacharya., A, (1997). Modelling Exponential Survival Data with Dependent Censoring, Sankya : 59, 242-267.

7. Breterler., et al. (2018). Reliability of Wireless Monitoring using a Wearable Patch Senson in high-risk Surgical Patients at a Step-down Unit in the Netherlands: a Clinical Validation Study. BMJ Open, Vol. 8, Issue. 2, pp. 1-10.

8. Bunea, F. and McKeague, I. W. (2005). Covariate Selection for Semiparametric Hazard Function Regression Models. Jour. of Multivariate Analysis, 92, 186-204.

9. Cai, J., et al. (2008). Partially linear hazard Regression with varying Co-efficients for Multivariate Survival Data. J.R.Statist., Series B, 70, part 1, 141-158.

10. Catchpole, E.A, et al. (2008). A new Method for Analysing discrete life History data with missing covariate values. J.R.Statist., Series B, 70, part 2, 445-460.