Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 1, pp. 1–8
DOI : 10.47974/JIOS-1291

Lasso estimation in parametric frailty model

Anu Sirohi *
Department of Mathematics
Sharda University
Greater Noida 201310
India

Prem Shanker Jha †
Department of Statistics
Patna College
Patna University
Bihar 800005
India

Abstract

This paper proposed lasso estimator in parametric frailty model. Comparison of lasso (least absolute shrinkage and selection operator) and maximum likelihood (ML) estimator is done in terms of scalar mean square error (MSE). Performance of lasso estimator is examined through simulation study. Furthermore, approach is applied to analyze infant mortality in India.

Subject Classification: *62NXX.*

Keywords: *Frailty model, Collinearity, Maximum likelihood estimator, Lasso, Child mortality.*

Introduction

Frailty model is an approach to deal with time to event data which consider observable and unobservable factors simultaneously. Time to event is defined as the time of the occurring of the event of interest (Rahardja and Wu 2018; Andersen et. al. 2008). But it has the assumption that the explanatory variables are independent to each other while in real life situations it rarely happens. If this assumption is not satisfied, ML estimator gives unstable results and variance matrix inflates Weissfeld

* *E-mail:* **anu.sirohi@sharda.ac.in (Corresponding Author)**
† *E-mail:* **premshankarjhabhu@gmail.com**

© Taru Publications

and Sereika (1991). Therefore alternatives of ML estimator are proposed to cover the problem of collinearity.

Tibshirani (1996) introduced a new estimator i.e. lasso in linear regression model. Further Tibshirani (1997) applied this approach in Cox proportional hazard regression model. This study applied lasso approach in parametric frailty model.

Lasso estimator for parametric frailty model

Frailty model is a famous model of survival analysis where event time is related to measured and unmeasured covariates at the same time. Mathematical formulation of the model is defined as,

$$\xi(t_i \mid x_i, \mu_i) = \mu_i \xi_0(t_i)\ \exp(\alpha' x_i) \tag{1}$$

where $\xi(t_i \mid x_i, \mu_i)$ represents the hazard of i^{th} individual at time t_i, $i = 1,2, \ldots, n$ with explanatory variables $x_i' = (x_{i1}, x_{i2}, \ldots, x_{iq})$ and unobserved frailty variable μ_i. $\xi_0(t_i)$ is the base line hazard function and $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_q)$ is the vector of regression coefficients. Here, q represents the number of regressors.

Sastry (1996) and Oakes (1982) assumed that the frailty variable follow gamma distribution with variance p^2. In the same manner this paper also assumed that the frailty variable follow the same distribution with variance p^2. Present study apply the parametric case of frailty model therefore it has the assumption that the outcome variable follows Weibull distribution (one parameter i.e. γ) and the shape of baseline hazard will be $\xi_0(t_i) = \gamma t_i^{\gamma-1}, \gamma > 0.$ Likelihood function of model (1) is given below,

$$L(\alpha) = \prod_{i=1}^{n} (\mu_i \xi_0(t_i) e^{\alpha' x_i})^{\delta_i}\ e^{\mu_i Z_0(t_i) e^{\alpha' x_i}} \tag{2}$$

where $Z_0(t_i)$ is the cumulative baseline hazard and δ_i is the censoring indicator. On taking the assumption of gamma distribution for frailty and on performing some algebraic calculation the likelihood function will be,

$$L(\alpha) = \prod_{i=1}^{n} \left(\frac{\xi_0(t_i) e^{\alpha' x_i}}{1 + p^2 Z_0(t_i) e^{\alpha' x_i}} \right)^{\delta_i} (1 + p^2 Z_0(t_i) e^{\alpha' x_i})^{-1/p^2} \tag{3}$$

On taking log in Equation (3)

$$\begin{aligned} l(\alpha) = \textstyle\sum_{i=1}^{n} [\delta_i [\log(\xi_0(t_i)) + \alpha' x_i - \log(1 + p^2 Z_0(t_i) e^{\alpha' x_i})] \\ -(1/p^2)\log(1 + p^2 Z_0(t_i) e^{\alpha' x_i}) \end{aligned} \tag{4}$$

On applying the Newton Raphson method in Equation (4), we will get the MLE $\hat{\alpha}$ for α. On this basis m^{th} approximation is,

$$\hat{\alpha}^{(m)} = \hat{\alpha}^{(m-1)} + S(\hat{\alpha}^{(m-1)})^{-1} T(\hat{\alpha}^{(m-1)}) \tag{5}$$

where, $T(\hat{\alpha}^{(m-1)})$ represents the differentiation of $l(\alpha)$ with respect to α using $\hat{\alpha}^{(m-1)}$,

$$T(\hat{\alpha}^{(m-1)}) = \sum_{i=1}^{n} [\delta_i x_i - \{(1+\delta_i p^2) Z_0(t_i) x_i e^{\alpha'^{(m-1)} x_i}\} / \{1 + p^2 Z_0(t_i) e^{\alpha'^{(m-1)} x_i}\}] \tag{6}$$

and $S(\hat{\alpha}^{(m-1)})$ represents the negative of the second differentiation of $l(\alpha)$ with respect to α using $\hat{\alpha}^{(m-1)}$,

$$S(\hat{\alpha}^{(m-1)}) = X'UX \tag{7}$$

where, U is the diagonal matrix of order $n * n$ where $(i, i)^{th}$ entry is,

$$u_{ii} = t_i^{\gamma} e^{\alpha'^{(m-1)} x_i} (1 + \delta_i p^2) / (1 + p^2 t_i^{\gamma} e^{\alpha'^{(m-1)} x_i})$$

To overcome the problem collinearity, this paper maximizes the penalized likelihood i.e.

$$l^k(\alpha) = l(\alpha) - k \mid \alpha \mid_1 \tag{8}$$

where $\mid \alpha \mid_1 = (\sum_{i=1}^{n} \mid \alpha \mid_1)$ is the L1 norm and k is the tuning parameter. On applying Newton Raphson method m^{th} approximation for α^k (lasso estimator) is,

$$\hat{\alpha}^{k(m)} = \hat{\alpha}^{k(m-1)} + S^k(\hat{\alpha}^{k(m-1)})^{-1} T^k(\hat{\alpha}^{k(m-1)}) \tag{9}$$

where,

$$T^k(\hat{\alpha}^{k(m-1)}) = T(\hat{\alpha}^{(m-1)}) - k\, sgn(\hat{\alpha}^{k(m-1)})$$

$$S^k(\hat{\alpha}^{k(m-1)}) = S(\hat{\alpha}^{(m-1)}) + k\lambda$$

where, $\lambda = diag\ (sgn'(\hat{\alpha}^{k(m-1)})$ and $sgn'(.)$ is the differentiation of $sgn(.)$. Asymptotic bias of α^k is,

$$E(\hat{\alpha}^k - \alpha) = -k(X'UX + k\lambda)^{-1} \alpha$$

and asymptotic variance is,

$$(X'UX + k\lambda)^{-1} X'UX (X'UX + k\lambda)^{-1}$$

In the same manner, asymptotic MSE is calculated by taking the sum of variance and square of biasness. The most common choices of k for ridge estimator in linear, logistic and Cox regression model have been $1/\widehat{\alpha'\hat{\alpha}}$, $q/\widehat{\alpha'\hat{\alpha}}$ and $(q+1)/\widehat{\alpha'\hat{\alpha}}$ (Schaefer et. al. 1984; Smith et. al. 1991; Xue et. al. 2007).

Simulation Study

To examine the performance of lasso and ML estimator simulation is conducted and compared the estimators in terms of MSE. Survival function (SF) is considered as,

$$S(t_i) = \exp(Z_0(t_i)\mu_i e^{\alpha' x_i})$$

and cumulative distribution function (CDF) is,

$$F(t_i) = 1 - \exp(Z_0(t_i)\mu_i e^{\alpha' x_i})$$

If CDF follow uniform distribution with range 0 to 1 then SF also follow the same distribution with the same range i.e. $V \sim U(0,1)$ then $1-V \sim U(0,1)$.

$$V = \exp(Z_0(t_i)\mu_i e^{\alpha' x_i}) \sim U(0,1)$$

and

$$t = Z_0^{-1}(-\log(V)/\mu_i e^{\alpha' x_i})$$

with the assumption of Weibull distribution $Z_0^{-1}(\ ) = t^{1/\gamma}$ now the failure time is,

$$t = (-\log(V)/\mu_i e^{\alpha' x_i})^{1/\gamma} \tag{10}$$

On the basis of (10), failure times are produced and size of the sample is taken as 50. Frailty variable is produced through gamma distribution with parameters 2 (shape) and 0.5 (scale). Two explanatory variables are generated in such a manner that they are correlated to each other and four categories ($\rho = 0.4,\backslash\ 0.6,\backslash\ 0.8$ and 0.95) of correlation are considered. Censoring times are produced through uniform distribution in the range 0 to 4.5 which have the censoring rate of 39 percent. One hundred data sets are generated and averaged MSE corresponding to each covariate are calculated for both estimators.

Table 1, represents the results of simulation which indicates that the lasso estimator has lesser MSE as compared to ML estimator. When

$\rho = 0.8$ then reduction in the percentage of MSE is increased from 66.38% to 85.93% in respect of the change from $k = 1/\widehat{\alpha'\hat{\alpha}}$ to $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$ respectively. For each category of ρ MSE is minimum for $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$. Reduction in the percentage of MSE is increased as ρ increased.

Real Life Application

This study utilized the data from the third phase of National Family Health Survey (NFHS III) India. 11581 infants are included out of which 10448 infants are censored. Outcome variable is the time of survival (in months) of infants. Three explanatory variables i.e. TT(Tetanus Toxoid), IFA (Iron Folic Acid) are ANC (Antenatal Care) are included, out of which TT and ANC are moderately correlated to each other.

Lot of research is available to examine the infant mortality (Pebley et. al. 1996; Singh et. al. 2007; Omariba et. al. 2007; Sirohi and Rai 2019a; Sirohi and Rai 2019b; Rai and Sirohi 2020; Hussain and Borah 2020; Pal and Pal 2021; Ahmmed et. al. 2020). But collinearity is seldom discussed in infant mortality. This paper discussed collinearity in frailty model. Proposed method is applied to this data set and results are given in Table 2. ML estimates says that if no tetanus injection is given during pregnancy then risk of infant death is significantly 1.352 times high as compared to the reference category. Risk of infant death is significantly decreased by 0.925 times, if mother was taking IFA during pregnancy. Estimated p^2 is significantly non zero that means unmeasured variables also affect the survival of infant. MSE of real data corresponding to both estimators are given in Table 3, which shows that MSE of lasso estimator is smaller as compared to MLE and it is minimum for $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$.

Conclusion

In this study lasso estimation is applied in parametric frailty which is an alternative of ML estimation in the presence of collinearity. Simulation Results represented that lasso generalized more précised and accurate estimates as compared to ML estimates. This comparison is done on the basis of MSE.

Future Scope

In this method lasso is applied on parametric frailty model. In future the method can be applied in non parametric frailty model.

Table 1

MSE comparison in simulation among the MLE and ridge estimators

ρ	Estimator	MSE (α)	Reduction in MSE (α)%
0.4	MLE	0.07998	
	α^k *for* $k = 1/\widehat{\alpha'\hat{\alpha}}$	0.05258	34.25
	α^k *for* $k = q/\widehat{\alpha'\hat{\alpha}}$	0.03869	51.62
	α^k *for* $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$	0.03037	62.02
0.6	MLE	0.10537	
	α^k *for* $k = 1/\widehat{\alpha'\hat{\alpha}}$	0.05778	45.16
	α^k *for* $k = q/\widehat{\alpha'\hat{\alpha}}$	0.03928	62.72
	α^k *for* $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$	0.02968	71.83
0.8	MLE	0.18809	
	α^k *for* $k = 1/\widehat{\alpha'\hat{\alpha}}$	0.06323	66.38
	α^k *for* $k = q/\widehat{\alpha'\hat{\alpha}}$	0.03698	80.33
	α^k *for* $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$	0.02646	85.93
0.95	MLE	0.68790	
	α^k *for* $k = 1/\widehat{\alpha'\hat{\alpha}}$	0.04886	92.89
	α^k *for* $k = q/\widehat{\alpha'\hat{\alpha}}$	0.02590	96.23
	α^k *for* $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$	0.01917	97.21

Table 2

Estimation in NFHS data

Category				
	exp $(\hat{\alpha})$	S.E. $(\hat{\alpha})$	exp $(\hat{\alpha}^k)$	S.E. $(\hat{\alpha}^k)$
TT				
No	1.352*	0.143	1.356*	0.132
1	1.062*	0.153	1.063*	0.150

Contd...

>=2[a]	-	-	-	-
ANC				
No	1.214*	0.151	1.211*	0.149
1-2	1.152*	0.108	1.169*	0.107
>=3[a]	-	-	-	-
IFA				
No[a]	-	-	-	-
Yes	0.925*	0.123	0.934*	0.119

a: Reference Category

*: Statistical Significance at 5% level of significance

S.E.: Standard Error

Table 3

MSE (NFHS data)

Method	**MSE (α)**	**Reduction in MSE (α)%**
MLE	4.552	
α^k *for* $k = 1/\widehat{\alpha'\hat{\alpha}}$	4.215	07.99
α^k *for* $k = q/\widehat{\alpha'\hat{\alpha}}$	3.120	25.98
α^k *for* $k = (q+1)/\widehat{\alpha'\hat{\alpha}}$	2.043	51.53

References

[1] Andersen, H., Splied, H., Larsen, S. And Dall, V. (2008). Statistical analysis of time to event data from preclinical safety pharmacology studies. *Journal of Statistics and Management Systems,* 10(2), 111-125.

[2] Ahmmed, M.M., Babu, M.A. and Ferdosy, J.C. (2020). Direct and indirect effects of COVID 19 on maternal and child health in Bangladesh. *Journal of Statistics and Management Systems,* 24(1), 175-192.

[3] Hussain, Z. And Borah, M.D. (2020). Birth Weight prediction of new born baby with application of machine learning techniques on feature of mother. *Journal of Statistics and Management Systems*, 23(6), 1079-1091.

[4] Oakes, D. (1982). A model for association in bivariate survival data. *Journal of Royal Statistical Society, (B)*, 44(3), 414-422.

[5] Omariba, D., Beaujot, R. and Rajulton, F. (2007). Determinants of infant and child mortality in Kenya: an analysis controlling for frailty effects. *Popul Res Policy Rev*, 26, 299-321.

[6] Pebley, A. R., Goldman, N., and Rodriguez, G. (1996). Prenatal and delivery care and childhood immunization in Guatemala: Do family and community matter? *Demography*, 33(2), 231-247.

[7] Pal, S.K. and Pal, A.K. (2021). The impact of increase in COVID 19 cases with expectational situation to SDG: Good health and wellbeing. *Journal of Statistics and Management Systems*, 24(1), 209-228.

[8] Rahardja, D. And Wu, H. (2018). Statistical methodological review for time to event data. *Journal of Statistics and Management Systems,* 21(1), 189-199.

[9] Rai, P. K. and Sirohi, A. (2020). On the behavior of Child Mortality through different Survival Models in India. *International Journal of Ecological Economics & Statistics*, 41(2), 19-33.

[10] Sastry, N. (1996). Community characteristics, individual and household attributes, and child survival in Brazil. *Demography,* 33(2), 211-229.

[11] Singh, K. K., Pandey, N. and Gautam A. (2007). Effect of Breastfeeding and Maternal Health Care programme on infant mortality. *Demography India*, 36(2), 253-266.

[12] Sirohi, A. and Rai, P. K. (2019a). On the Structure of Infant Mortalitv using Accelerated Failurc Time (AFT) Model: A Comparative Study based on National Family Health Survey (NFHS) Data in India. *Indian Journal of Public Health Research and Development*, 10(2), 135-143.

[13] Sirohi, A. and Rai, P. K. (2019b). An Analytical Framework for the Determinants of Infant Mortality based on 2005-06 NFHS Data in India. *Journal of Ecophysiology & Occupational Health*, 19(1\&2), 37-47.

[14] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, 58, 267-288.

[15] Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16, 385-395.

[16] Weissfeld, L. A. and Sereika, S. M. (1991). A multicollinearity diagnostic for generalized linear models *Communications in Statistics - Theory and Methods*, 20(4), 1183–1198.