Journal of Statistical Modelling: Theory and Applications
Vol. 2, No. 1, 2021, pp. 143-150
Yazd University Press 2021

Research Paper

Introduction to shared frailty Cox models with parametric and non-parametric distributions and their application in medical data

Navideh Nikmohammadi[1,2], Parvin Sarbakhsh[2],
S. Morteza Shamshirgaran[3], Neda Gilani*[2]
[1]Student Research Committee, Tabriz University of Medical Sciences,
Tabriz, Iran
[2]Department of Statistics and Epidemiology,
Tabriz University of Medical Science, Tabriz, Iran
[3]Department of Statistics and Epidemiology,
Faculty of Medical Sciences Neyshabur, Neyshabur, Iran

Received: October 19, 2020/ Revised: February 23, 2021 / Accepted: April 24, 2021

Abstract: In survival data, it is typical for survival times to be clustered or depend on some unobserved covariates. This can be due to geographical clustering, subjects sharing common genes, specific socioeconomic level, or hereditary and racial characteristics, and other predisposition that cannot be measured and observed directly. Adjusting the effects of these unknown factors on the survival functions is necessary for the independence of survival times and the explanatory variables. The aim of this study is to introduce and compare Cox models with parametric and non-parametric shared frailty on brain stroke survival data. The results showed that non-parametric frailty model has better fitting than parametric distributions (AIC=4686 and BIC=4684), especially when the exact parametric distribution is not known. According to the results of best model, following variables were statistical significant; BMI (HR=0.97, P=0.045); Age (HR=1.04, P<0.001); HDL (HR=1.01, P<0.001); LDL (HR=0.99, P<0.001); Hyperlipidemia (HR=0.72, P<0.014). The nonparametric frailty is desirable, due to potential misspecification of the parametric form and as a method for detecting clusters of groups with similar frailties.

Keywords: Cox model, Shard frailty, Non-Parametric distribution, Brain stroke.
Mathematics Subject Classification (2010): 62N01, 62G08, 62P10, 62H30.

*Corresponding author: `gilanin@tbzmed.ac.ir`

1 Introduction

The semi-parametric Cox proportional hazards (PH) regression model was developed by David Cox (1972) and is by far the most favored model for survival analysis. The partial likelihood (Cox 1975) and Breslow (1974) estimator were used to estimate parameters in the cox model. In the survival models, it is assumed that the time of events are independent of each other, but there are situations where this assumption does not hold due to the existence of some unknown and unmeasured variables which are related to the study event. Most of the studies have considered the frailty term as a parameter in the Cox model, with the specified distribution, such as Hougaard et al. (a,b 1986), Austin et al. (2017), and Hougourd et al. (1995). These models have optimal fitting if the parametric distribution is specified correctly, otherwise, the model estimate will have potential biases. Gasperoni et al. (2020) introduced a semi-parametric Cox model with non-parametric frailty.

A non-parametric alternative is desirable for the distribution of frailty due to potential misspecification of the parametric form and as a method for detecting clusters of groups with similar frailty. This process extends the shared frailty Cox model to include frailty that has a separate distribution with an unknown number of elements in its support. Thus no defined structure is imposed on either the clustering or the baseline survival. This structure creates both a very flexible model and a probabilistic clustering technique, which is used to explore heterogeneity between groups. Also, this model is appropriate for analyzing the large and categorized from multicentral data.

2 Frailty term

The dispersion between time-to-event data often occurs due to errors and omissions of other effective variables. The difference between the survival of individuals (shorter or longer individual survival than other individuals) not considered with conventional models. Walper et al. (1979) introduced the frailty and applied the effects of unmeasured elements and invisible heterogeneity in the survival models. The random effect of frailty is like a random variable on the regression models. In the univariate frailty model effect of each person is independent than others and the hazard and survival function is written as follows.

$$h\left(t|\alpha\right) = \alpha h(t), \tag{1}$$

$$S\left(t|\alpha\right) = S\left(t\right)^{\alpha}. \tag{2}$$

2.1 Shared frailty

In the shared frailty model, it is assumed that the frailty of individuals in the groups is identical and independent from other groups. Common frailty is a way to calculate the correlation in data which is due to invisible factors that are common within each group. The general form of the hazard and survival function with common frailty is defined as follows:

$$h_{ij}\left(t\right) = h_0\left(t\right) u_i \exp\left(X_{ij}^t \beta\right); i = 1, \dots s \quad j = 1 \dots n_i, \tag{3}$$

$$S_{ij}\left(t\right)=\exp\left(-H_0\left(t\right)u_i\exp\left(X_{ij}^t\beta\right)\right);i=1,\ldots s\quad j=1\ldots n_i, \tag{4}$$

where $h_0\left(t\right)$ represents the baseline hazard, β is the vector of regression coefficients, and u_i is the frailty term of i group, X_{ij}^t vector of covariates, $H_0\left(t\right)$ is cumulative hazard function.

2.2 Semi-parametric Cox model with parametric shared frailty

In this model, the baseline hazard function has non-parametric distribution, but the predictor component and shared frailty term have parametric distribution. Parametric distribution can be gamma, log-Normal, inverse Gaussian and etc. Estimation of parametric component is performed by expectation-maximization (EM) algorithm. In the following model, the likelihood function does not have closed form due to the baseline hazard, which is unknown. To solve this problem and obtain sufficient information to estimate the parameters of β vector, a partial likelihood function was proposed by Cox (1972).

$$L=\prod_{i=1}^{n}\left[(1-G(y_i))f(y_i)\right]^{\delta_i}\left[\left(1-F\left(y_i\right)\right)g\left(y_i\right)\right]^{1-\delta_i}. \tag{5}$$

We introduce g and G, as notation for the density function and cumulative distribution function of the censoring time.

For right-censored data, the actual information for subject i, $i=1,...,n$, is contained in the pair (y_i,δ_i), where y_i is the minimum of the event time t_i and the censoring time $c_i, y_i=\min(t_i,c_i)$ and δ_i is the censoring indicator, taking the value one if the event has been observed, otherwise δ_i takes value zero.

If the censors are non-informative, then the likelihood function is as follows;

$$L=\prod_{i=1}^{n}\left(f\left(y_i\right)\right)^{\delta_i}\left(S(y_i)\right)^{1-\delta_i}=\prod_{i=1}^{n}\left(h(y_i)\right)^{\delta_i}S(y_i). \tag{6}$$

2.3 Semi-parametric Cox model with a non-parametric shared frailty

Shared frailty with a non-parametric distribution in Cox model is another model for estimation of heterogeneity of survival in groups of time-to-event data. It's generalization of the Cox model. The frailty term has separate distribution by an unknown number of points in its support. This model detects clusters of groups with similar frailties. Its assumption that each statistical unit belongs to one group.

$$h\left(t:X_{ij},\widetilde{w}_k,z_{jk}\right)=\prod_{k=1}^{K}\left[\lambda_0\left(t\right)\widetilde{w}_k\exp\left(X_{ij}^T\beta\right)\right]^{z_{jk}} \tag{7}$$

where λ_0 represents the baseline hazard, β is the vector of regression coefficients, and $\widetilde{w}_k$ is the shared frailty among groups of the same latent population k. Both the frailty and the baseline hazard are assumed to be non-parametric. The model (7) is an extension of a proportional hazard Cox model. The observable data Y are made up of the set of $Y_{ij}=\{T_{ij},\delta_{ij},X_{ij}\}$ over all i,j. We define this as the "incomplete" data, while the “complete” data are the realizations of the vector $\{T_{ij}, \delta_{ij}, X_{ij}, w_k, z_{ij}\}$, we also assume that censoring is non-informative, thus that T^*_{ij} and C_{ij} are conditionally independent, given X_{ij}, $\widetilde{w}_k$ and T_{jk}. Selection of the number of clusters is based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the approach of Laird (1978). For estimation parameters in this method was used from steps of expectation-maximization (EM) algorithm. See Gasperoni et al. (2020) for more information. Semi-parametric

Cox model with non-parametric shared frailty by having a discrete frailty distribution, together with an unspecified baseline hazard, leads to a relatively novel and very flexible model for grouped survival data, so don’t have potential misspecification of the parametric form. It’s a new flexible model for detecting heterogeneity of data by selection clusters with similar frailty. The motivation of this work is computational efficiency investigated that they would have a significant effect on the analysis of very large databases, such as the administrative clinical database. Such administrative multi-central databases are powerful tools for finding questions in epidemiology and other medical research. This model has been used for the survival of patients with heart failure by Agosti Michela et al. (2018), also the study of the effect of hospitals on the research on death rate on lung cancer patients by Federico Rea et al. (2020).

3 An application study to comparison of models on the brain stroke patients

The data information is from 1306 patients with first-ever brain stroke conducted during a prospective cohort study from November 2013 to March 2017 in Emam Reza and Razi hospitals of Tabriz city. Patients with first-ever ischemic and hemorrhagic stroke, specified by the International Classification of Diseases (ICD-10) system with final diagnosis based on computed tomography (CT) and magnetic resonance imaging (MRI) scan, were included and followed for two years. To measure the degree of disability in patients with stroke, the Modified Rankin Scale (MRS) was used. The written consent was taken from patients or their proxies, and information was completed by an instructed expert. The study event was death after the first-ever stroke in patients. Also, the survival time was time to death in these patients, by Nubakht et al. (2020). In the present study, we used the shared frailty models for analysis of the brain stroke survival. The Emam Reza hospital is better known than Razi and, it admitted most of its patients from small cities in East Azerbaijan Province and a few of Tabriz city, versus the Razi hospital admitted its patients from the city of Tabriz itself. Also, each of the hospitals has specific treatment policies, such as the number of patients per nurse, the workload of medical centers staff, medical centers staff’s specialization, or other hidden reasons that can influence the survival of the patients who refer to each center with specific characteristics, social and economic conditions, etc. Health care providers in the form of hospitals or health centers play an important role in the survival and improvement of stroke patients. So, estimation the quality of health services provided by the health centers is important to health care systems for reassessment survival prognostic and risk stratification Hong-Sheng Du et al. (2016).

4 Result

A total of 1036 patients were studied. The median follow-up was 730 days. The rate of mortality after the first ever stroke was 38 percent during this follow-up period. The mean (± SD) of age at diagnosis patients was 69.07(±12.79) years. The mean (± SD) of the body mass index (BMI) (kg/m2) of patients was 25.79 (±4.54). The mean (± SD) of hospitalized and laboratory tests for patients were hyperlipidemia: 0.47 mg/dl (±0.49), HDL: 43.92 mg/dl (±13.79), LDL: 106.52 mg/dl (±48.42). The complete details have been reported in Table 1.

Table 1: Descriptive statistics of brain stroke patients

Variables	Minimum	Maximum	Mean	Standard Deviation
BMI (kg/m2)	15.43	55.56	25.79	4.54
Age (year)	15.00	94.00	69.07	12.79
HDL (mg/dl)	15.00	171.00	43.92	13.79
LDL (mg/dl)	24.00	698.00	106.52	48.42
HPL (mg/dl)	0.00	1.00	0.47	0.49

BMI=Body mass index, HDL=High-density lipoprotein,
LDL=Low-Density Lipoprotein, HPL= high blood fats.

Represented Cox models with shared frailty of gamma, inverse Gaussian, and non-parametric distribution, respectively. The finding of Cox model with non-parametric frailty showed that every five covariates: BMI, age, HDL, LDL, Hyperlipidemia were statistical significant factors in survival rate. In the models with parametric frailty term, with gamma distribution HDL, LDL, Hyperlipidemia, and in the inverse Gaussian model; Age, HDL, Hyperlipidemia had statistically significant results. The model with non-parametric frailty, in addition to measuring the effect of variables on the hazard function, had identified two latent populations among the dataset (table3). AIC and BIC criteria for models with non-parametric frailty term were AIC=4686.00, and BIC=4684.00. Also, for Cox model with gamma frailty were AIC=5837.64 and BIC=5835.69; and for Cox model with Inverse-Gaussian frailty were AIC=5830.88 and BIC=5828.93. The conclusion of these criteria showed that the Cox model with non-parametric frailty had better fitting than other models. According to the best model, significant variables are: BMI (HR=0.97, P=0.045); Age (HR=1.04, P<0.001); HDL (HR=1.01, P<0.001); LDL (HR=0.99, P<0.001); Hyperlipidemia (HR=0.72, P<0.014), the complete details have been reported in Table 2.

In the internal best fit of the Cox model with non-parametric frailty, the best fit for the number of latent populations, according to AIC, BIC and Laird criteria was estimated to be 2 latent populations. The dependence of the observations on each of these latent populations is $\frac{1}{2}$. The frailty ratio among populations is estimated to be approximately 4.73, (have been showed in Table 3 and Figure 1).

Table 2: Semi-parametric Cox models with parametric and nonparametric shared frailty

	Frailty distribution								
	Gamma			Invers-Gaussian			Non-parametric		
Variables	HR¥	SE†	P	HR¥	SE†	P	HR¥	SE†	P*
BMI (Kg/m2)	0.97	1.02	0.297	0.97	1.02	0.253	0.97	1.01	0.045*
Age (year)	1.03	1.03	0.237	1.03	1.01	0.045*	1.04	1	<0.001*
HDL (mg/dl)	1.01	1.00	0.003*	1.01	1.003	<0.001*	1.01	1	<0.001*
LDL (mg/dl)	0.99	1.00	0.045*	0.99	1.002	0.133	0.99	1	<0.001*
HPL (mg/dl)	0.76	1.12	0.022*	0.76	1.12	0.02*	0.72	1.14	0.014*

BMI=Body mass index, HDL=High-density lipoprotein, LDL=Low-Density Lipoprotein, HPL= high blood fats.

¥ Hazard ratio
† Standard error
∗ Significant at 0.05 level

Table 3: Selection of the number latent population in cox model with Non-parametric frailty

Population	Estimates	Std. Err*	Log-lik®	AIC	BIC	Optimal K		
P1	0.5	0.353	-2394.67	4801.35	4825.21	Laird	AIC	BIC
P2	0.5	0.353	-2337.89	4691.78	4723.59	2	2	2
w2/w1	4.738	0.512						

BIC: Bayesian information criterion, AIC: Akaike information criterion.

® log-likelihood
∗ Standard error

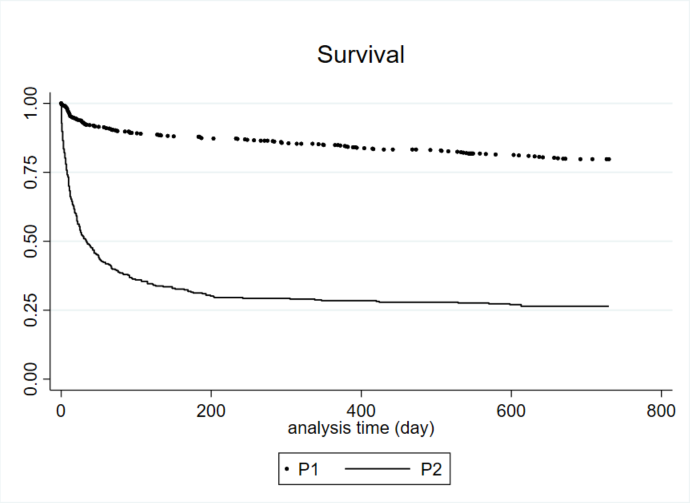

Figure 1: Semi-parametric cox model with non-parametric frailty for two cluster

5 Conclusions

Time-to-event data analysis is widely used in medicine. Survival models with frailty terms are the conventional method for analyzing cluster data in multiple health-care centers. Frailty is a random term of unknown and invisible characteristics that can influence the results of survival data that enter the model as a multiplicative factor with a specific distribution. Parametric survival models with parametric frailty (gamma, log-logistic, log-normal, Weibull, and etc.) were used for analyzing these data. And semi-parametric Cox model with parametric and non-parametric frailty term. In the study, we compared the semi-parametric Cox model with parametric and non-parametric distribution for shared frailty terms. The results showed a model with non-parametric frailty had better fitting than the parametric form of frailty. In the model with non-parametric frailty estimated two latent populations; members of each population have the same hazard and placed in each population by probability equal $\frac{1}{2}$. If the appropriate distribution is chosen for the frailty term, then that model will be better than models with non-parametric frailty term. But when the exact distribution for frailty is not specified, the model with a nonparametric frailty term will fit well. In this model, it's not necessary to specify the distribution of the baseline hazard and strong parametric assumptions for the frailty term; Potential biases due to the placement of the parametric form of the frailty not included. It can also be considered as a possible clustering method to investigate the heterogeneity of survival at the cluster level. Clinically, this model can be used to analyze data categorized with specific characteristics in hospitals, nursing homes, research centers, and groups of patients with specific genetics and treatment responses. Also, in health centers; repeated evaluations for each patient can be used.

Acknowledgments

This article is taken from a student dissertation. The tracking and ethics code is 64214 and IR.TBZMED.REC.1398.832, respectively. It has been done in partnership with Tabriz University of Medical Sciences and Emam Reza and Razi hospitals of Tabriz city.

References

Agosti, M. (2018). Applications of nonparametric frailty models for the analysis of long term survival in heart failure patients, http://hdl.handle.net/10589/140100.

Austin, P.C. (2017). A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review*, **85**(2), 185-203.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187-202.

Duchateau, L. and Janssen, P. (2007). *The Frailty Model*, Springer Science & Business Media.

Du, H.S., Ma, J.J. and Li, M. (2016). High-quality health information provision for stroke patients. *Chinese Medical Journal*, **129**(17), 2115.

Fine, J.P., Glidden, D.V. and Lee, K.E. (2003). A simple estimator for a shared frailty regression model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(1), 317-329.

Gasperoni, F., Ieva, F., Paganoni, A.M., Jackson, C.H. and Sharples, L. (2020). Non-parametric frailty Cox models for hierarchical time-to-event data. *Biostatistics*, **21**(3), 531-544.

Hougaard, P. (1986a). A class of multivanate failure time distributions. *Biometrika*, **73**(3), 671-678.

Hougaard, P. (1986b). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**(2), 387-396.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, **1**(3), 255-273.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**(364), 805-811.

Lemeshko, B.Y., Lemeshko, S.B., Akushkina, K.A., Nikulin, M. S. and Saaidia, N. (2010). Inverse Gaussian model and its applications in reliability and survival analysis. In *Mathematical and statistical models and methods in reliability* (pp. 433-453). Birkhäuser, Boston, MA.

Novbakht, H., Shamshirgaran, S.M., Sarbakhsh, P., Savadi-Oskouei, D., Yazdchi, M.M. and Ghorbani, Z. (2020). Predictors of long-term mortality after first-ever stroke. *Journal of Education and Health Promotion*, **9**(45), PMC7161659.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*, Stata Press.

Rea, F., Ieva, F., Pastorino, U., Apolone, G., Barni, S., Merlino, L. and Corrao, G. (2020). Number of lung resections performed and long-term mortality rates of patients after lung cancer surgery: evidence from an Italian investigation. *European Journal of Cardio-Thoracic Surgery*, **58**(1), 70-77.

Vaupel, J.W., Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**(3), 439-454.

Wienke, A. (2010). *Frailty Models in Survival Analysis.* CRC press.