Journal of Engineering and Applied Sciences 14 (19): 7069-7071, 2019
ISSN: 1816-949X

Rank Estimators Versus Least Square Estimators for Estimating the Parameters of Semiparametric Accelerated Failure Time Model

Mostafa Karimi and Noor Akma Ibrahim
Institute for Mathematical Research, University Putra Malaysia, Sari Kembangan, Malaysia
mostafa.karimi.ir@gmail.com, +601139864131

Abstract: Rank-based method and least square approach are the most common techniques for estimating the regression parameters of accelerated failure time model. In this study, both inference procedures are considered their advantages and disadvantages are explained and their similarities and differences are discussed.

Key words: Accelerated failure time model, rank-based inference, least square method, semiparametric method, censored data, linear regression, biostatistics

INTRODUCTION

Accelerated failure time model is an appealing regression model to biostatistics researchers due to its simple interpretation (Karimi and Shariat, 2017). Estimating the regression parameters of the model through parametric methods is quite challenging in the presence of censored observations (Karimi *et al*., 2017). In such cases, semiparametric approaches are very common. Two main semiparametric methods for estimating the unknown parameters of the model are rank-based method (Jin *et al*., 2013) and least square method (Jin *et al*., 2006). In this study, both inference procedures are briefly explained and their main theoretical and computational aspects are considered. Both approaches are also compared and their advantages and disadvantages are discussed. The main focus of this study is on investigating the similarities and differences of two methods in theory as well as their performance in applications.

MATERIALS AND METHODS

Inference procedures

Accelerated failure time model: For the ith subject of a random sample of n subjects let T_i denote the failure time, C_i denote the censoring time and Z_i denote the p×1vector of corresponding covariates. Assume that conditional on covariates Z_i, failure times T_i and censoring times C_i are independent. The accelerated failure time model takes the form:

$$\text{Log}(T_i) = \beta' Z_i + \epsilon_i \quad (1)$$

where, β is a p-vector of unknown model parameters and ϵ_i are the error terms of the model for i = 1, ..., n with a common distribution function F which is unspecified (Kalbfleisch and Prentice 2011). The data consists of $(\tilde{T}_i, \delta_i, Z_i)$ where $\tilde{T}_i$ is the minimum of T_i and C_i and δ_i is 1, if $T_i \le C_i$ and 0 otherwise. The introduced model is a semiparametric linear regression model which relates the log-transformed failure times to the covariates.

Rank estimators: Define $\tilde{e}_i(b) = \text{Log}(\tilde{T}_i) - b'Z_i, \tilde{Y}_i(t;b) = I[\tilde{e}_i(b) \ge t]$ where, $I\{.\}$ is the indicator function and $S^{(a)}(t;b) = n^{-1}\sum_{i=1}^{n}\tilde{Y}_i(t;b)(Z_i)^a$ for a = 0, 1. The weighted log-rank estimating function for the unknown parameter β is given by:

$$U_\phi(b) = \sum_{i=1}^{n}\delta_i\phi(\tilde{e}_i(b),b)\{Z_i - \bar{Z}(\tilde{e}_i(b),b)\} \quad (2)$$

where, $\bar{Z}(\tilde{e}_i(b),b) = S^{(1)}(\tilde{e}_i(b);(b))/S^{(0)}(\tilde{e}_i(b);b)$ and $\phi(\tilde{e}_i(b),b)$ is a weight function. The estimating function correspond to Gehan (1965), if $\phi(\tilde{e}_i(b),b) = S^{(0)}(\tilde{e}_i(b);(b))$ and log-rank if $\phi(\tilde{e}_i(b),b) = 1$ (Mantel, 1966). Let $\hat{\beta}_R$ denote, the rank estimator for the unknown parameter of the model which is the solution of $\{U_\phi(b) = 0\}$. For estimating the unknown parameters of the model Jin *et al.* (2013) proposed an iterative algorithm on the basis of the general weighted estimating function. The algorithm at its kth iteration is given by:

$$\hat{\beta}_R^{(k)} = \arg\min_b L_R\left(b, \hat{\beta}_R^{(k-1)}\right) \quad (3)$$

Where:

$$L_R\left(b,\hat{\beta}\right) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\phi\left(\tilde{e}_i\left(\hat{\beta}\right),\hat{\beta}\right)/ S^{(0)}\left(\tilde{e}_i\left(\hat{\beta}\right);\hat{\beta}\right)\{\tilde{e}_j(b) - \tilde{e}_i(b)\}I\{\tilde{e}_j(b) \ge \tilde{e}_i(b)\} \quad (4)$$

Corresponding Author: Mostafa Karimi, Institute for Mathematical Research, University Putra Malaysia, Sari Kembangan, Malaysia, mostafa.karimi.ir@gmail.com, +601139864131

According to Jin *et al.* (2013), the rank estimator $\tilde{\beta}_R^{(k)}$ is asymptotically normal for any k.

Least square estimators: When there is no censored observations the least square estimator of the unknown model parameters is obtained by solving the following estimating Eq. 1:

$$\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)\left(\text{Log}\left(T_i\right)-b'Z_i\right)=0 \quad (5)$$

This estimating equation cannot be used when data contains censored observations, since, the actual value of T_i is unknown for subject i when $\delta_i = 0$. For obtaining the least square estimators in the presence of censored data (Jin *et al.*, 2006) proposed an iterative algorithm which at its kth iteration is given by:

$$\tilde{\beta}_R^{(k)}=L_s\left(\hat{\beta}_s^{(k-1)}\right) \quad (6)$$

Where:

$$L_s(b)=\left\{\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)\left(Z_i-\bar{Z}\right)'\right\}^{-1}\left\{\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)\left(\hat{Y}_i(b)-\bar{Y}(b)\right)'\right\} \quad (7)$$

In Eq. 7: $\hat{Y}_i(b)=n^{-1}\sum_{i=1}^{n}\hat{Y}_i(b)$ and $\hat{Y}_i(b)=E\left(\text{Log}\left(T_i\right)\middle|\tilde{T}_i,\delta_i,Z_i\right)$ which is proposed by Buckley and James (1979) and can be approximated by:

$$\hat{Y}_i(b)=\delta_i\text{Log}\left(\tilde{T}_i\right)+(1-\delta_i)\left\{\frac{\int_{\tilde{e}_i(b)}^{\infty}u\,d\hat{F}(u)}{1-\hat{F}\left(\tilde{e}_i(b)\right)}+b'Z_i\right\} \quad (8)$$

where, $\hat{F}$ is the Kaplan-Meier estimator of F. According to Jin *et al.* (2006), the least square estimator $\hat{\beta}_s^{(k)}$ is asymptotically normal, if the initial value $\hat{\beta}_s^{(0)}$ is asymptotically normal.

RESULTS AND DISCUSSION

Both the rank-based method and the least square approach are semiparametric inference procedures, since, the probability distribution of error terms of the model is completely unknown. One advantage of rank-based inference over the least square method is that it does not involve estimating the distribution of the error terms while obtaining least square estimators requires the Kaplan-Meier estimator of the distribution of the error terms. This makes the least square method and its corresponding algorithm more complicated than the rank-based method, both theoretically and computationally. Note that both algorithms need a consistent estimator of the model parameter such as Gehan estimator for their initial values. Thus, the least square approach requires to obtain a rank estimator prior to the computational stage of its associated algorithm. In addition, it has been established that rank estimators are always asymptotically normal (Tsiatis, 1990; Ying, 1993) while the asymptotic normality of least square estimators strongly depend on the asymptotic normality of the initial value of their corresponding algorithm. However, the results of the simulation studies by Jin *et al.* (2006) illustrated that there was no significant difference between the efficiency of rank estimators and least square estimators. More precisely, the rank estimators were slightly more efficient under extreme-value error and the least square estimators were slightly more efficient under logistic and normal erors.

CONCLUSION

For estimating the regression parameters of semiparametric accelerated failure time model both rank estimators and least square estimators are common. From a theoretical point of view, rank-based inference procedure involves less technical difficulties, since, it does not require estimating the probability distribution of the error terms while least square approach involves Kaplan-Meier estimator of the distribution of the error terms. Moreover, the asymptotic normality of rank estimators does not depend on the distribution of the initial value of its associated algorithm. In application, the results of simulation studies show that there is no significant difference between the efficiency of rank estimators and least square estimators. Therefore, in studies that researcher is free to choose between these two methods rank estimators are definitely more recommended than least square estimators.

REFERENCES

Buckley, J. and I. James, 1979. Linear regression with censored data. Biometrika, 66: 429-436.

Gehan, E.A., 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika, 52: 203-223.

Jin, Z., D.Y. Lin and Z. Ying, 2006. On least-squares regression with censored data. Biometrika, 93: 147-161.

Jin, Z., D.Y. Lin, L.J. Wei and Z. Ying, 2013. Rank-based inference for the accelerated failure time model. Biometrika, 90: 341-353.

Kalbfleisch, J.D. and R.L. Prentice, 2011. The Statistical Analysis of Failure Time Data. 2nd Edn., Wily-Interscience, New York, USA., Pages: 435.

Karimi, M. and A. Shariat, 2017. Semiparametric accelerated failure time model as a new approach for health science studies. Iran. J. Public Health, 46: 1594-1595.

Karimi, M., N.A. Ibrahim, M.R.A. Bakar and J. Arasan, 2017. Rank-based inference for the accelerated failure time model in the presence of interval censored data. Numer. Algebra, Control Optim., 7: 107-112.

Mantel, N., 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother. Rep., 50: 163-170.

Tsiatis, A.A., 1990. Estimating regression parameters using linear rank tests for censored data. Annal. Stat., 18: 354-372.

Ying, Z., 1993. A large sample study of rank estimation for censored regression data. Annal. Stat., 21: 76-99.