RANK-BASED INFERENCE FOR THE ACCELERATED FAILURE TIME MODEL IN THE PRESENCE OF INTERVAL CENSORED DATA

Mostafa Karimi*, Noor Akma Ibrahim, Mohd Rizam Abu Bakar and Jayanthi Arasan
Institute for Mathematical Research, Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Abstract. Semiparametric analysis and rank-based inference for the accelerated failure time model are complicated in the presence of interval censored data. The main difficulty with the existing rank-based methods is that they involve estimating functions with the possibility of multiple roots. In this paper a class of asymptotically normal rank estimators is developed which can be acquired via linear programming for estimating the parameters of the model, and a two-step iterative algorithm is introduced for solving the estimating equations. The proposed inference procedures are assessed through a real example. The results of applying the proposed methodology on the breast cancer data show that the algorithm converges after three iterations, and the estimations of model parameter based on Log-rank and Gehan weight functions are fairly close with small standard errors.

1. Introduction. In failure time data analysis the accelerated failure time model is quite attractive to researchers as an alternative to the popular Cox proportional hazard model [4], due to its easy physical interpretation. This model relates the covariates linearly to the logarithm of the failure time [9].

For the semiparametric analysis of accelerated failure time model, a class of linear rank estimators was proposed by [16] based on the weighted log-rank statistics. In the early 1990’s [17], [18], [20], [10], [11], and [22] studied the asymptotic properties of the rank estimators, among others.

To estimate the parameters of the model, [21] suggested a sieve maximum likelihood estimator as a semiparametric method with application to interval censored HIV data. Semiparametric estimation methods were also considered by [8], [15], [19], and [23], while the rank-based inference was recently studied by [3], [11], [14], and [2].

Semiparametric methods for the accelerated failure time model have had significant theoretical advances, but they have rarely been used in applications. The main difficulty with the existing computational method is that they are not reliable

2010 Mathematics Subject Classification. Primary: 62N01, 62N02; Secondary: 62G09, 90-08.
Key words and phrases. Survival analysis, semiparametric analysis, rank-based inference, accelerated failure time model, interval censored data.

* Corresponding author: M. Karimi.
The reviewing process of the paper was handled by Wah June Leong as Guest Editor.
and efficient in practice. To be precise, the existing rank-based estimating functions possibly result multiple roots, since they are step functions of the model parameters.

In this paper, simple and reliable methods are provided for estimating the parameters of the aforementioned accelerated failure time model in the presence of interval censored data. The proposed methodology modifies a generalisation of the rank estimators previously studied by [20], [18], and [22]. To obtain the rank estimator with the Gehan-type weight function, it is shown that the minimization of the Gehan’s estimating function can be carried out by minimizing a convex objective function. A class of general weighted log-rank estimating functions is introduced to approximate the log-rank estimating functions. These estimating equations are solved through an iterative algorithm. The proposed two-step algorithm includes estimation step and approximation step, which can be executed through linear programming in each iteration.

2. Inference procedures.

2.1. Model. Let random variable $T_i$ be the failure time for the $i$th subject for $i = 1, 2, ..., n$, and let $Z_i$ be a $p \times 1$ associated vector of covariates. Assuming that the failure times are independent, the accelerated failure time model specifies that

\[
\log T_i = \beta_0^T Z_i + \varepsilon_i \quad (i = 1, ..., n),
\]

where $\beta_0$ is a $p \times 1$ vector of unknown regression parameters and $\varepsilon_i$’s are error terms which are independent and identically distributed, but their distribution is unspecified.

Suppose that the failure time for subject $i$, $T_i$ cannot be observed directly, but it can be examined at random examination times. For the $i$th subject let $R_i = (R_{i1}, R_{i2}, ..., R_{in_i})$ denote the sequence of ordered examination times, where $n_i$ is the number of examinations. Assume that examination times and failure times are independent. For the subject $i$, let $R_{iL}$ be the last examination time prior to $T_i$, and let $R_{iU}$ be the first examination time posterior to $T_i$. It is assumed that the true unknown $T_i$ lies in the interval $(R_{iL}, R_{iU})$.

Let $C_i$ denote the censoring time for the $i$th subject. It is assumed that $C_i$ are independent of $T_i$ conditionally on $Z_i$. The data then consist of $(R_{iL}, \tilde{R}_{iU}, \Delta_i, Z_i)$, where $\tilde{R}_{iL} = R_{iL} \wedge C_i$, $\tilde{R}_{iU} = R_{iU} \wedge C_i$, and $\Delta_i = I\{R_{iU} \leq C_i\}$ ($a \wedge b = \min(a, b)$ and $I\{\cdot\}$ is the indicator function).

Let random variable $\tilde{T}_i$ indicates an approximation of the unknown true failure time $T_i$, which satisfies the assumption $R_{iL} < \tilde{T}_i \leq R_{iU}$. Define $N^+_i(t; \beta) = \Delta_i I\{e_i^*(\beta) \leq t\}$ and $Y_i^*(t; \beta) = I\{e_i^*(\beta) \geq t\}$, where $e_i^*(\beta) = \log \tilde{T}_i - \beta^T Z_i$. Write

\[
S^{(0)}(t; \beta) = n^{-1} \sum_{i=1}^{n} Y_i^*(t; \beta), \quad S^{(1)}(t; \beta) = n^{-1} \sum_{i=1}^{n} Y_i^*(t; \beta) Z_i.
\]

The proposed weighted log-rank estimating function for $\beta_0$ is as follows:

\[
U_\phi(\beta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t; \beta) \{Z_i - \tilde{Z}(t; \beta)\} dN_i^+(t; \beta)
\]

or

\[
U_\phi(\beta) = \sum_{i=1}^{n} \Delta_i \phi(e_i^*(\beta); \beta) \{Z_i - \tilde{Z}(e_i^*(\beta); \beta)\}
\]
where \( \bar{Z}(t; \beta) = S^{(1)}(t; \beta)/S^{(0)}(t; \beta) \), and \( \phi \) is a specified weight function which is twice continuously differentiable on \([0, 1]\). Note that \( U_\phi(\beta) \) is the log-rank estimating function \([13]\) if \( \phi(t; \beta) = 1 \), and it is the Gehan estimating function \([17]\) if \( \phi(t; \beta) = S^{(0)}(t; \beta) \). A reasonable estimator for \( \beta_0 \) is a value of \( \beta \), denoted by \( \hat{\beta}_0 \), which is a zero-crossing of \( U_\phi(\beta) \) or a minimiser of \( ||U_\phi(\beta)|| \). The random vector \( n^{1/2}(\hat{\beta}_0 - \beta_0) \) is asymptotically zero-mean normal, according to the general asymptotic theory for the rank estimators \([18]; [10]; [22]\).

However, it is not easy to solve the system of equations \( U_\phi(\beta) = 0 \), since \( U_\phi(\beta) \) is in general a \( p \)-dimensional step function of \( \beta \). In fact, \( U_\phi(\beta) \) is non-monotone and discontinues in \( \beta \), thus it is difficult to solve the equation or locate a minimiser.

2.2. Proposed methodology. The Gehan estimating function can be written as follows:

\[
U_G(\beta) = \sum_{i=1}^{n} \Delta_i S^{(0)}(e_i^*(\beta); \beta)\{Z_i - \bar{Z}(e_i^*(\beta); \beta)\}
\]

or

\[
U_G(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) I(e_i^*(\beta) \leq e_j^*(\beta)) \tag{3}
\]

which is monotone in each component of \( \beta \) \([6]\). Considering that \( (Z_i - Z_j) I(e_i^*(\beta) \leq e_j^*(\beta)) \) is the gradient of \( \{e_i^*(\beta) - e_j^*(\beta)\}^- \) in \( \beta \) (since \( a^- = |a|I\{a < 0\} \)), it is not difficult to see that the Gehan estimating function \( U_G(\beta) \) is the gradient of the loss function

\[
L_G(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (e_i^*(\beta) - e_j^*(\beta))^-. 
\]

The minimizer of \( L_G(\beta) \), denoted by \( \hat{\beta}_G \), is a root of the Gehan estimating function. As considered by \([12]\), this minimization can be carried out through minimizing the linear function \( \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i u_{ij} \), where \( u_{ij} \geq 0 \) and \( u_{ij} \geq -\{e_i^*(\beta) - e_j^*(\beta)\}^- \). The random vector \( n^{1/2}(\hat{\beta}_G - \beta_0) \) is asymptotically zero-mean normal according to the general asymptotic theory for the rank estimators, as mentioned before.

To construct weighted Gehan-type loss functions which their minimisers can be obtained by linear programming, an approach similar to that of \([8]\) will be developed. Consider the following modification of the weighted log-rank estimating function

\[
\bar{U}_\phi(\hat{\beta}; \beta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \psi(t + (\hat{\beta} - \hat{\beta})'Z_i) S^{(0)}(t; \beta)\{Z_i - \bar{Z}(t; \beta)\} dN_i^*(t; \beta) \tag{4}
\]

where \( \psi(t; \beta) = \phi(t, \beta)/S^{(0)}(t; \beta) \) and \( \hat{\beta} \) is a preliminary estimator of \( \beta_0 \), such as \( \hat{\beta}_G \). Clearly,

\[
\bar{U}_\phi(\hat{\beta}_G; \beta) = \sum_{i=1}^{n} \psi \{e_i^*(\beta) + (\hat{\beta}_G - \beta)^'Z_i, \hat{\beta}_G\} S^{(0)}(e_i^*(\beta); \beta)\{Z_i - \bar{Z}(e_i^*(\beta); \beta)\}
\]

\[
= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(e_i^*(\hat{\beta}_G), \hat{\beta}_G) \Delta_i (Z_i - Z_j) I(e_i^*(\beta) \leq e_j^*(\beta)) .
\]

Note that \( \bar{U}_\phi(\hat{\beta}; \beta) \) is the same as \( U_G(\beta) \), except for \( \psi(e_i^*(\hat{\beta}), \hat{\beta}) \) which are free of \( \beta \). Thus, monotone in each component of \( \beta \), \( \bar{U}_\phi(\hat{\beta}; \beta) \) is the gradient of the following
loss function

\[ \hat{L}_\phi(\hat{\beta}; \beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(e_i^*(\hat{\beta}), \hat{\beta}) \Delta_i \{e_i^*(\beta) - e_i^*(\beta)\}^- . \]

Similar to the case of \( L_G(\beta) \), the minimisation of \( \hat{L}_\phi(\hat{\beta}; \beta) \) is a linear optimization problem.

For inference in the presence of interval censored data, we will propose a two-step iterative algorithm, which estimates \( \beta_0 \) at the first step and approximates the censored observations at the second step. Due to the fact that the interval censored failure times are unknown, the estimating function in general can be written as \( U_\phi(\beta; \{\hat{T}_i; i = 1, ..., n\}) \). The proposed estimation-approximation algorithm at the \( k \)th iteration can be described in two steps as follows:

**Estimation step:** Determine \( \hat{\beta}^{(k)}_\phi \) as the the minimizer of

\[ ||U_\phi(\beta; \{\hat{T}_i^{(k-1)}; i = 1, ..., n\})||, \]

or as the root of \( U_\phi(\beta; \{\hat{T}_i^{(k-1)}; i = 1, ..., n\}) \).

**Approximation step:** Find the \( \{\hat{T}_i^{(k)}\} \) that \( R_iL < \hat{T}_i^{(k)} \leq R_iU \) for \( i = 1, ..., n \), and maximizes \( U_\phi(\hat{\beta}^{(k)}_\phi; \{\hat{T}_i; i = 1, ..., n\}) \).

The algorithm starts from an initial value of \( \{\hat{T}_i^{(0)}\} \), and cycles back and forth between the estimation and approximation steps. However, the steps of the algorithm can also be written on the basis of the general weighted loss function at the \( k \)th iteration as follows:

**Estimation step:** Determine \( \hat{\beta}^{(r)}_\phi \) as the minimizer of

\[ \hat{L}_\phi(\beta; \hat{\beta}^{(r-1)}_\phi; \{\hat{T}_i^{(r-1)}; i = 1, ..., n\}) \]

**Approximation step:** Find the \( \{\hat{T}_i^{(k)}\} \) that \( R_iL < \hat{T}_i^{(k)} \leq R_iU \) for \( i = 1, ..., n \), and maximizes

\[ \hat{L}_\phi(\hat{\beta}^{(k)}_\phi; \hat{\beta}^{(k-1)}_\phi; \{\hat{T}_i; i = 1, ..., n\}) \]

where \( \hat{\beta}^{(0)}_\phi = \hat{\beta}_G \). It is important to point out that the convergence of the estimators is not fundamental to the validity of the proposed method, because for any \( k \), \( \hat{\beta}^{(k)}_\phi \) and \( \{\hat{T}_i^{(k)}\} \) themselves are legitimate estimators of \( \beta_0 \) and \( \{\hat{T}_i\} \), respectively.

### 3. Numerical results.

#### 3.1. Breast cancer data.

The proposed method is illustrated with the data which were obtained from a retrospective study on early breast cancer for 94 patients, reported by [5]. The data consist of 46 patients who were given radiation therapy alone (RT), and 48 patients who were given radiation therapy plus adjuvant chemotherapy (RTC). At visits, which were supposed to be at clinic every 4 to 6 months, the breast retraction of the patients were evaluated by physicians. However, the actual visit times and times between visits differ from patient to patient. Whereas the exact time to breast retraction was not observed, the data contain the observed time periods during which breast retractions occurred, represented by intervals in months.

To determine the effect of the treatments on the time to breast retractions, the treatment indicator \( Z_i \) was considered as the covariate in the accelerated failure time model (model (1)). To be clear, \( Z_i \) represents RT group and RCT group by
the values 0 and 1, respectively. The estimation-approximation algorithm was used to obtain the estimates based on the log-rank and the Gehan weight functions. The estimates were identical after five iterations, with the accuracy of 0.0001 between successive estimates. The confidence intervals were constructed using the Wald method. The main results of the regression parameter estimates are summarized in Table 1.

To find the failure times that maximize the estimating function in the approximation step of the algorithm the behaviour of Log-rank and Gehan estimating functions were examined using the presented failure time data of 94 patients. Figure 1 displays the line charts for Log-rank and Gehan estimating functions versus the failure times of 10th and 30th patients which are members of the RT group, and 60th and 90th patients which are members of the RTC group. As illustrated in Figure 1, for patients in the same group of therapy methods, the behaviour of estimating functions were very similar. To be specific, the estimating functions increase when the failure times increase. Therefore, maximizing the estimating function in the approximation step can be carried out by taking the upper bound of the censoring interval as the approximation of the failure time for each subject.

4. Conclusion. The introduced iterative algorithm is defined based on a class of general weighted log-rank estimating functions and their corresponding loss functions, and it does not involve estimating functions that are step functions of the model parameters. The first step of the algorithm is computationally simple, since it requires solving the estimating equations which are convex functions of the model parameters. Although the maximization part in the second step of the algorithm may create difficulties, the results of the numerical study displays that examining the behaviour of the estimating function for the possible values of approximated failure times helps to avoid complications in practical use.
REFERENCES

[1] M. Bouadoumou, Y. Zhao and Y. Lu, Jackknife empirical likelihood for the accelerated failure time model with censored data, *Communications in Statistics-Simulation and Computation*, 44 (2014), 1818–1832.

[2] S. H. Chiou, S. Kang and J. Yan, Rank-based estimating equations with general weight for accelerated failure time models: an induced smoothing approach, *Statistics in Medicine*, 34 (2015), 1495–1510.

[3] M. Chung, Q. Long and B. A. Johnson, A tutorial on rank-based coefficient estimation for censored data in small- and large-scale problems, *Statistics and computing*, 23 (2013), 601–614.

[4] D. R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society, 34* (1972), 187–220.

[5] D. M. Finkelstein and R. A. Wolfe, A semiparametric model for regression analysis of interval-censored failure time data, *Biometrics*, 41 (1985), 933–945.

[6] M. Fygenson and Y. Ritov, Monotone estimating equations for censored data, *The Annals of Statistics*, 22 (1994), 732–746.

[7] E. A. Gehan, A generalized Wilcoxon test for comparing arbitrarily single-censored samples, *Biometrika*, 52 (1965), 203–223.

[8] Z. Jin, D. Y. Lin, L. J. Wei and Z. Ying, Rank-based inference for the accelerated failure time model, *Biometrika*, 90 (2003), 341–353.

[9] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.

[10] T. L. Lai and Z. Ying, Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, *The Annals of Statistics*, 19 (1991a), 1370–1402.

[11] T. L. Lai and Z. Ying, Rank regression methods for left-truncated and right-censored data, *The Annals of Statistics*, 19 (1991b), 531–556.

[12] D. Y. Lin, L. J. Wei and Z. Ying, Accelerated failure time models for counting processes, *Biometrika*, 85 (1998), 605–618.

[13] N. Mantel, Evaluation of survival data and two new rank order statistics arising in its considerations, *Cancer Chemotherapy Reports*, 50 (1966), 163–170.

[14] J. Ning, J. Qin and Y. Shen, Semiparametric accelerated failure time model for length-biased data with application to dementia study, *Statistica Sinica*, 24 (2014), 313–333.

[15] L. Peng and J. P. Fine, Regression modeling of semicompeting risks data, *Biometrics*, 63 (2007), 96–108.

[16] R. L. Prentice, Linear rank tests with right censored data, *Biometrika*, 65 (1978), 167–179.

[17] Y. Ritov, Estimation in a linear regression model with censored data, *The Annals of Statistics*, 18 (1990), 303–328.

[18] A. A. Tsiatis, Estimating regression parameters using linear rank tests for censored data, *The Annals of Statistics*, 18 (1990), 354–372.

[19] Y. G. Wang and L. Fu, Rank regression for accelerated failure time model with clustered and censored data, *Computational Statistics and Data Analysis*, 55 (2011), 2334–2343.

[20] L. J. Wei, Z. Ying and D. Y. Lin, Linear regression analysis of censored survival data based on rank tests, *Biometrika*, 77 (1990), 845–851.

[21] H. Xue, K. F. Lam, B. J. Cowling and F. de Wolf, Semi-parametric accelerated failure time regression analysis with application to interval-censored HIV/AIDS data, *Statistics in medicine*, 25 (2006), 3850–3863.

[22] Z. Ying, A large sample study of rank estimation for censored regression data, *The Annals of Statistics*, 21 (1993), 76–99.

[23] J. Zhang and Y. Peng, Semiparametric estimation methods for the accelerated failure time mixture cure model, *Journal of the Korean Statistical Society*, 41 (2012), 415–422.

Received April 2016; 1st revision May 2016; final revision November 2016.