Regression analysis of interval censored competing risk data using a pseudo-value approach

Sooyeon Kim\textsuperscript{a}, Yang-Jin Kim\textsuperscript{1,a}

\textsuperscript{a}Department of Statistics, Sookmyung Women's University, Korea

Abstract

Interval censored data often occur in an observational study where the subject is followed periodically. Instead of observing an exact failure time, two inspection times that include it are available. There are several methods to analyze interval censored failure time data (Sun, 2006). However, in the presence of competing risks, few methods have been suggested to estimate covariate effect on interval censored competing risk data. A subdistribution hazard model is a commonly used regression model because it has one-to-one correspondence with a cumulative incidence function. Alternatively, Klein and Andersen (2005) proposed a pseudo-value approach that directly uses the cumulative incidence function. In this paper, we consider an extension of the pseudo-value approach into the interval censored data to estimate regression coefficients. The pseudo-values generated from the estimated cumulative incidence function then become response variables in a generalized estimating equation. Simulation studies show that the suggested method performs well in several situations and an HIV-AIDS cohort study is analyzed as a real data example.

Keywords: competing risks, cumulative incidence function, GEE, interval censored data, pseudo-value approach

1. Introduction

A competing risk model has been applied to provide more valid statistical inferences when the subjects are at risk of failure from several causes (Geskus, 2015). For investigating the effect of covariate on the cause-related failure, two approaches have been applied: the cause-specific hazard (CSH) and the subdistribution hazard (SH). The CSH measures the instantaneous hazard of a particular cause-related failure by regarding other causes-related failures as censoring. The SH has a direct relation with a cumulative incidence function (CIF) that also make it possible to extend the result in order to interpret the covariate effect on the CIF (Fine and Gray, 1999). However, the SH utilizes a distribution of right censoring variable to reflect a possible contribution of competing failures to risk set. In a context of interval censored failure time, the distribution of such a censoring variable seems complex. For example, under a mixed case interval censored data, a right censoring is defined as the last inspection time of having no failure and should therefore be specified in the inspection process (Schick and Yu, 2000). Therefore, the direct application of Fine and Gray’s method to interval censored data seems to be inappropriate.

Andersen et al. (2003) and Klein and Andersen (2005) presented a pseudo-value approach to directly model the effect of covariate on the cumulative incidence function. The objective is to replace...
the true but unknown survival times with pseudo-values. If we want to estimate \( \theta = E(f(t)) \) for a function \( f \) of failure time \( t \), then it can be calculated with \( 1/n \sum_{i=1}^{n} f(t_i) \) with a known \( t_i \). For example, \( f(t) = I(t > u) \) becomes a survival function. Now, pseudo-values are defined as

\[
\hat{\theta}_i = n \hat{\theta} - (n - 1) \hat{\theta}_{i-1},
\]

where \( \hat{\theta}_{i-1} \) is the estimator \( \hat{\theta} \) calculated based on the sample leaving out the \( i^{th} \) observation. Our interest is to estimate covariate effect on failure time with censoring. In more detail, define \((T_i, \delta_i), i = 1, \ldots, n\), where \( T_i = T_{i1} \wedge C_i \) and \( \delta_i = I(T_i = T_{i1}) \) with a censoring time \( C_i \) and a failure time \( T_{i1} \). Denote \( Z_i \) as a covariate vector of subject \( i \) and \( D_i \) as a cause indicator, respectively. Assume that the censoring time \( C_i \) is independent of \((T_i, D_i, Z_i)\), where \( D_i = \delta_i D_{i1} \). Define \( N_{ik}(t) = I(T_{i1} \leq t, D_i = k) \) as the count of the \( k^{th} \) cause failures over \([0, t]\) and \( Y_i(t) = I(T_{i1} \geq t) \) as at-risk indicator of subject \( i \), respectively. Then, the cumulative incidence function is estimated as the simple version of the Aalen-Johansen estimator, that is, a special case of a multi-state model with a transition probability \( P_{ik}(s, t) \) and is defined as

\[
\hat{F}_k(t) = \int_0^t \hat{S}(u-)d\hat{A}_k(u) = \hat{P}_{1k}(0, t),
\]

where

\[
\hat{A}_k(t) = \int_0^t \frac{\sum_{i=1}^{n} dN_{ik}(u)}{\sum_{i=1}^{n} Y_i(u)}
\]

is the Nelson-Aalen estimator of the cumulative cause specific hazard function and \( \hat{S}(t) \) is the Kaplan-Meier estimator calculated using all causes failures. Klein and Andersen (2005) proposed the use of pseudo-values to estimate a covariate effect on the cumulative incidence function. These pseudo-values are calculated at the grids of time-points \( w_1 < w_2 < \cdots < w_s \) chosen from observed failure times. Too many grids give a burden to calculate too many values; therefore, five to ten time-points equally spaced on the failure time scale have been suggested. Graw et al. (2009) showed the unbiasedness using a second-order von Mises expansion and Jacobsen and Martinussen (2016) utilized a U-statistic to establish the estimating equation and proposed a new type of variance.

Some studies have been proposed for an interval censored competing risk data. For a current status data, Jewell et al. (2003) developed a NPMLE and a pseudo MLE of a CIF and Jewell and Kalbfleisch (2004) suggested a modified pool adjacent violator algorithm (PAVA). For general interval censored data, Hudgens et al. (2001) suggested NPMLEs. However, only a few works have been done on a regression model. Sun and Shen (2009) proposed a two-stage estimation procedure for a cause specific event incidence probability and a hazard function. Hudgens et al. (2014) suggested a parametric regression model by extending Jeong and Fine (2006)’s method.

In this paper, we consider an extension of a pseudo-value approach to interval censored data. In Section 2, the estimation procedure is described and the performance of the suggested method is evaluated via several simulation studies in Section 3. Section 4 presents the result of the application to a real dataset and some discussions are commented on in Section 5.

2. Statistical model

Denote \( l_i, r_i, d_i, z_i \) as the observable data set from a subject \( i = 1, \ldots, n \). Assume that censoring times \( l_i \) and \( r_i \) satisfying \( l_i < t_i < r_i \) are independent of a failure time \( t_i \). The first step of a pseudo-value approach is to estimate CIF. By extending Turnbull’s (1974) self-consistency algorithm, Hudgens et al.
Regression analysis of interval censored competing risk data using a pseudo-value approach

(2001) suggested two versions of nonparametric estimators. The NPMLE of CIF has a cause specific time support that results in undefined regions when implementing with overall survival functions. As an alternative approach, a pseudolikelihood estimator (PLE) was calculated with a pooled time support derived from all causes interval censored data. In this paper, we adopt the PLE to estimate CIF for interval censored data. Denote \( E = \bigcup_{i=1}^{m} [q_i, p_i] \) as an equivalence set which is a set of nonoverlapping intervals. \( [q_i, p_i] \) is defined as the elements of left censoring times and the ones of right censoring times following immediately (Lindsay and Ryan, 1998; Peto, 1973). Now, denote \( a^k_d \) as an indicator variable, \( a^k_d = I[[q_i, p_i] \in [l_i, r_i] \cap d_i = k) \) and let \( \psi^k_I = F_k(p_i^+) - F_k(q_i^-) \). Then a pseudolikelihood is defined as \( \prod_{i=1}^{m} [F_d(r_i^-) - F_d(l_i^-)]^q S(l_i^{1-k}) \). Based on Hudgens’ Lemmas (Hudgens et al., 2001) related with \( E \) and \( F_k \), this likelihood is rewritten as

\[
\prod_{i=1}^{n} \prod_{l=k}^{n} \prod_{j=1}^{m} [a^k_d \psi^k_I].
\]

To estimate \( \psi = \psi^k_I \), define an indicator \( l^k_d = \{ t_i \in [q_i, p_i] \cap d_i = k \} \). However, this quantity is unavailable owing to unknown failure times \( t_i \). To solve this problem, the EM algorithm (Dempster et al., 1977) is applied. In E-step, given the data \( O_i = l_i, r_i, d_i, \) and \( \psi \),

\[
E \left( \psi^k_I | \psi, O_i \right) = \psi^k_I \psi \frac{\psi^k_I}{\sum_{k=1}^{K} \psi^k_I} = \psi^k_I \psi \frac{\sum_{k=1}^{K} \psi^k_I}{\sum_{k=1}^{K} \sum_{l=1}^{m} a^k_d \psi^k_I}
\]

is calculated. Then at M-step, the proportion of failure of cause \( k \) at the \( l^k_d \) interval is obtained with

\[
\psi^k_I(\psi) = \sum_{i=1}^{n} \psi_i \frac{\psi^k_I}{\sum_{l} \psi^k_I}, \quad \psi^k_I = \sum_{k=1}^{K} \psi^k_I.
\]

With initial values of \( \{ \psi^k_I = 1/(Km), \quad k = 1, \ldots, K; l = 1, \ldots, m \} \), iterate (2.1) and (2.2) until a convergence criterion satisfies. Using the final values of \( \{ \psi^k_I \} \), the NPMLE of CIF is defined as \( \hat{F}_k(t) = \sum_{p=1}^{l} \psi^k_I \) if \( p < t < q_i+1 \) and \( \hat{F}_k(t) = \psi^k_I + \cdots + \psi^k_m \) if \( t > p_n \).

The next step after estimating a CIF is to generate the Jackknife pseudo-values at the prefixed grids of time points \( w_1 < \cdots < w_r \),

\[
\theta_{ij} = n \hat{F}_k(w_j) - (n-1) \hat{F}_k^{-1}(w_j), \quad j = 1, \ldots, r.
\]

In order to find grid points, the equivalence set is utilized. Define \( e_l = (q_i + p_i)/2, \quad l = 1, \ldots, m \). Then \( w_1 < \cdots < w_r \) are determined as the points with equal distances among the sorted \( e_l \)’s. By treating \( \theta_j = (\theta_{1j}, \ldots, \theta_{rj}) \) as the repeated response variables and applying a suitable link function \( g \), the following generalized estimating equation (Liang and Zeger, 1986) is applied

\[
U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \left( \frac{\partial g^{-1}(\beta' z_i)}{\partial \beta} \right)' V^{-1} \left[ \theta_i - g^{-1}(\beta' z_i) \right] = 0,
\]

where \( g(\theta_i) = (g(\theta_{1j}), \ldots, g(\theta_{rj}))' \), \( g(\theta_{ij}) = \alpha_j + \gamma z_i = \beta' z_{ij} \) with \( \beta = (\alpha_1, \ldots, \alpha_r, \gamma) \) and \( V_i = \text{Cov}(\theta_{ij}, \theta_{ij}) \) is a working covariance matrix where AR(1) and independence covariance are most commonly used. Graw et al. (2009) remarked the condition of an asymptotic unbiasedness and expressed
Table 1: Estimation of a binary covariate effects using pseudo-values

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>Bias</th>
<th>ESE</th>
<th>SSE</th>
<th>CP</th>
<th>Independence</th>
<th>$p$</th>
<th>Bias</th>
<th>ESE</th>
<th>SSE</th>
<th>CP</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>100</td>
<td>0.009</td>
<td>0.448</td>
<td>0.442</td>
<td>0.952</td>
<td></td>
<td>0.008</td>
<td>0.467</td>
<td>0.450</td>
<td>0.966</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.019</td>
<td>0.317</td>
<td>0.312</td>
<td>0.965</td>
<td></td>
<td>0.019</td>
<td>0.332</td>
<td>0.320</td>
<td>0.968</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.012</td>
<td>0.259</td>
<td>0.251</td>
<td>0.946</td>
<td></td>
<td>0.014</td>
<td>0.271</td>
<td>0.257</td>
<td>0.944</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>100</td>
<td>0.008</td>
<td>0.326</td>
<td>0.316</td>
<td>0.948</td>
<td></td>
<td>0.019</td>
<td>0.343</td>
<td>0.328</td>
<td>0.946</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.003</td>
<td>0.225</td>
<td>0.226</td>
<td>0.964</td>
<td></td>
<td>0.005</td>
<td>0.237</td>
<td>0.236</td>
<td>0.965</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.005</td>
<td>0.189</td>
<td>0.184</td>
<td>0.948</td>
<td></td>
<td>0.006</td>
<td>0.198</td>
<td>0.193</td>
<td>0.954</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Estimation of a continuous covariate effects using pseudo-values

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>Bias</th>
<th>ESE</th>
<th>SSE</th>
<th>CP</th>
<th>Independence</th>
<th>$p$</th>
<th>Bias</th>
<th>ESE</th>
<th>SSE</th>
<th>CP</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>100</td>
<td>0.020</td>
<td>0.645</td>
<td>0.638</td>
<td>0.943</td>
<td></td>
<td>0.023</td>
<td>0.670</td>
<td>0.648</td>
<td>0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.012</td>
<td>0.442</td>
<td>0.449</td>
<td>0.950</td>
<td></td>
<td>0.020</td>
<td>0.465</td>
<td>0.452</td>
<td>0.947</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.013</td>
<td>0.348</td>
<td>0.349</td>
<td>0.956</td>
<td></td>
<td>0.014</td>
<td>0.353</td>
<td>0.351</td>
<td>0.956</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>100</td>
<td>0.020</td>
<td>0.470</td>
<td>0.467</td>
<td>0.947</td>
<td></td>
<td>0.017</td>
<td>0.485</td>
<td>0.471</td>
<td>0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.009</td>
<td>0.320</td>
<td>0.310</td>
<td>0.950</td>
<td></td>
<td>0.020</td>
<td>0.334</td>
<td>0.324</td>
<td>0.943</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.014</td>
<td>0.250</td>
<td>0.249</td>
<td>0.956</td>
<td></td>
<td>0.017</td>
<td>0.260</td>
<td>0.257</td>
<td>0.960</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the Jackknife pseudo-values as the von Mises expansion to derive the asymptotic properties including the following sandwich variance.

$$\hat{\Sigma} = \Gamma^{-1}(\hat{\beta}) \text{Var}(U(\hat{\beta})) \Gamma^{-1}(\hat{\beta}),$$

where

$$\Gamma^{-1}(\hat{\beta}) = \sum_{i=1}^{n} \left( \frac{dg^{-1}(\beta z_i)}{d\beta} \right)^{'} V_i^{-1} \left( \frac{dg^{-1}(\beta z_i)}{d\beta} \right)_{|\beta=\hat{\beta}},$$

and

$$\text{Var}(U(\hat{\beta})) = \sum_{i=1}^{n} U_i(\beta) U_i(\beta)^{'}_{|\beta=\hat{\beta}}.$$

3. Simulation

In this section, the performance of our proposed method is demonstrated with some simulated data. 500 replications with sample size $n = 100, 200$ and $300$ are generated. We consider two types of covariates: a binary covariate $Z = \{0, 1\}$ generated from a Bernoulli distribution with $p = 0.5$ and a continuous covariate $Z \sim N(0, 0.4)$. For simplicity, two causes are assumed ($K = 2$). Then a cause 1 failure time is generated with the following cumulative incidence function

$$F_1(t|Z) = 1 - \exp(-\Lambda_0(t) \exp(yZ)),$$

where $\Lambda_0(t) = \int_{0}^{t} \lambda_0(s) ds$ with $\lambda_0(t) = pe^{-t}/(1 - p(1 - e^{-t}))$ and $p$ is the occurrence probability of cause 1 failure, that is, $F_1(\infty|Z = 0) = p$. In this simulation, two values of $p$ are used ($p = 0.3$ or $0.6$). In order to make interval censored data, the different number of inspections for each subject is generated from a discrete uniform $h_i \sim U(10, 20)$ and then $h_i$'s inspection gap times are generated
Regression analysis of interval censored competing risk data using a pseudo-value approach

Table 3: Results of estimating effects of covariates for subtype E based on Independence

<table>
<thead>
<tr>
<th>TGP</th>
<th>Covariate</th>
<th>logit model</th>
<th>cloglog</th>
<th>logit model</th>
<th>cloglog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>γ</td>
<td>se(γ)</td>
<td>p-value</td>
<td>γ</td>
</tr>
<tr>
<td>4</td>
<td>Age</td>
<td>-0.039</td>
<td>0.015</td>
<td>0.011</td>
<td>-0.036</td>
</tr>
<tr>
<td></td>
<td>Gender (female = 1)</td>
<td>0.782</td>
<td>0.358</td>
<td>0.028</td>
<td>0.719</td>
</tr>
<tr>
<td>5</td>
<td>Age</td>
<td>-0.035</td>
<td>0.015</td>
<td>0.023</td>
<td>-0.032</td>
</tr>
<tr>
<td></td>
<td>Gender (female = 1)</td>
<td>0.859</td>
<td>0.364</td>
<td>0.018</td>
<td>0.782</td>
</tr>
<tr>
<td>6</td>
<td>Age</td>
<td>-0.038</td>
<td>0.015</td>
<td>0.013</td>
<td>-0.035</td>
</tr>
<tr>
<td></td>
<td>Gender (female = 1)</td>
<td>0.831</td>
<td>0.358</td>
<td>0.020</td>
<td>0.759</td>
</tr>
</tbody>
</table>

Table 4: Results of estimating effects of covariates for subtype E based on AR

<table>
<thead>
<tr>
<th>TGP</th>
<th>Covariate</th>
<th>logit model</th>
<th>cloglog</th>
<th>logit model</th>
<th>cloglog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>γ</td>
<td>se(γ)</td>
<td>p-value</td>
<td>γ</td>
</tr>
<tr>
<td>4</td>
<td>Age</td>
<td>-0.037</td>
<td>0.015</td>
<td>0.014</td>
<td>-0.035</td>
</tr>
<tr>
<td></td>
<td>Gender (female = 1)</td>
<td>0.789</td>
<td>0.360</td>
<td>0.028</td>
<td>0.719</td>
</tr>
<tr>
<td>5</td>
<td>Age</td>
<td>-0.035</td>
<td>0.015</td>
<td>0.019</td>
<td>-0.032</td>
</tr>
<tr>
<td></td>
<td>Gender (female = 1)</td>
<td>0.872</td>
<td>0.371</td>
<td>0.018</td>
<td>0.789</td>
</tr>
<tr>
<td>6</td>
<td>Age</td>
<td>-0.035</td>
<td>0.015</td>
<td>0.016</td>
<td>-0.034</td>
</tr>
<tr>
<td></td>
<td>Gender (female = 1)</td>
<td>0.837</td>
<td>0.363</td>
<td>0.021</td>
<td>0.762</td>
</tr>
</tbody>
</table>

from $g_i \sim U(0.05, 0.1)$. Then set $(L_i, R_i)$ as an interval censored data when satisfying $L_i = a_{i-1} < T_i < R_i = a_i$ with $L_i = a_{i-1} = a_{i-2} + g_{i-1}, R_i = a_{i-1} + g_i$ and $a_0 = 0$. If $T_i > a_{h_i}$, $T_i$ is regarded as a right censored with $(L_i, R_i) = (a_h, \infty)$. Now, pseudo-values were calculated at 6 grid time points ($w_1 < \cdots < w_6$) selected from the equivalence sets described at the previous section.

Tables 1 and 2 show the simulation results for a binary covariate and a continuous covariate, respectively. Each table includes the absolute bias of $\hat{γ}$, empirical standard error (ESE), mean of sandwich standard error (SSE), and 95% coverage probabilities (CP) under two working covariance structures (independence and AR(1)). The estimates show that biases are small for all cases and standard errors calculated from the sandwich estimator and ESE are similar. Also, the empirical coverage probabilities almost satisfy nominal level. Compared results with two different $p$ values, a larger $p(= 0.6)$ value with more cause 1 failures results in smaller standard errors than a smaller $p(= 0.3)$ one. Also, compared two covariance structures, the ESE under AR(1) is larger than one under independence working covariance.

4. Data analysis

The suggested method is applied to a HIV vaccine study designed to investigate the rates of HIV incidence and determine related risk factors (Hudgens et al., 2002). This project was established to assess the feasibility of the vaccine to HIV in the injecting drug users (IDU) in Bangkok, Thailand. 1209 HIV seronegative IDU were enrolled and they were supposed to visit about every four months for counseling and assessment of HIV seroconversion. A total of 1124 people had at least one visit, with 133 diagnosed with HIV seroconversion among them. In this study, two subtypes strains such as subtype B strain and subtype E strain can occur and the occurrence of one subtype censors one of the other subtype. In detail, of the 133 converts, 27 and 99 subjects have subtype B and subtype E, respectively, and the remaining seven patients’ subtypes were unknown. We investigate the effect of age and gender (female = 1) on the subtype E with the suggested method. To generate pseudo-values, three different numbers of grids (TGP = 4, 5, 6) are implemented using equivalence sets and two link functions (logit and complementary log-log) are applied. Tables 3 and 4 show estimated covariate
effects under two different covariance structures. According to results, older patients have a lower incidence rate than younger ones and female patients have a higher incidence rate than male patients. The results are similar among two covariance structures, two link functions and three different numbers of time grids. Figure 1 shows the estimated CIFs of subtype E of two genders by applying the PLE described in Section 2. From it, female patients seem to have a higher occurrence rate than male patients during the whole time period. However, this result should be dealt carefully because there were only fourteen female patients among the 133 ones.

5. Discussion

As the most widely used competing risk regression model, Fine and Gray’s subdistribution hazard model is based on an inverse probability censoring weighting technique that requires a specification of the distribution of censoring times. However, for interval censored data with different censoring structure, such specification is neither amenable nor possible. Therefore, we suggest the extension of the pseudo-value regression model proposed by Klein and Andersen (2005) to interval censored data instead of applying the Fine and Gray’s model. Jackknife pseudo-values are obtained from the estimated cumulative incidence function. The suggested approach can be implemented with an ordinary program since all calculations are performed using R package pseudo once estimating the CIFs. Simulation results show the proposed method results in consistent estimates and desirable coverage probabilities. The suggested pseudo-value approach can be applied to a multi-state model with several transition states and corresponding probabilities. In the studies on a multi-state model, Barrett et al. (2011) developed an interval censored semi-competing risk model that allowed a one direction transition between terminal events. Kim (2014) analyzed a bivariate current status data using a multi-state model. Commenges (2002) provided a review of the multi-state model with interval censored data. Another possible future work is about a missing cause. The IDU dataset includes seven subjects with unidentified HIV subtypes. Even though the number of the subjects with a missing cause is small in this dataset, this problem has commonly occurred in competing risk data; therefore, suitable methods should be developed for interval censored data (Goetghebeur and Ryan, 1995; Lu and Tsiatis, 2001; Moreno-Betancur and Latouche, 2013).
Acknowledgements

This work was supported by the National Research Foundation of Korea research grant NRF-2014R1A2A2A01003567. We are grateful to Professor Michael G. Hudgens for providing with IDU cohort study dataset.

References


Received August 11, 2016; Revised November 11, 2016; Accepted November 20, 2016