Communications for Statistical Applications and Methods
2016, Vol. 23, No. 4, 335–341

http://dx.doi.org/10.5351/CSAM.2016.23.4.335
Print ISSN 2287-7843 / Online ISSN 2383-4757

A modified estimating equation for a binary time varying covariate with an interval censored changing time

Yang-Jin Kim[1,a]

[a]Department of Statistics, Sookmyung Women's University, Korea

Abstract

Interval censored failure time data often occurs in an observational study where a subject is followed periodically. Instead of observing an exact failure time, two inspection times that include it are made available. Several methods have been suggested to analyze interval censored failure time data (Sun, 2006). In this article, we are concerned with a binary time-varying covariate whose changing time is interval censored. A modified estimating equation is proposed by extending the approach suggested in the presence of a missing covariate. Based on simulation results, the proposed method shows a better performance than other simple imputation methods. ACTG 181 dataset were analyzed as a real example.

Keywords: binary time-varying covariate, changing time, estimating equation, interval censored, failure time data, missing covariate, imputation method

1. Introduction

There exist several approaches to analyze interval censored failure time data. However, only few methods have been suggested for an interval censored covariate. This paper proposes a method to analyze a failure time data with a binary time-varying covariate whose changing time is interval censored. In an ordinary survival analysis. For example, well-known heart transplantation data includes a covariate about whether a subject received a heart transplantation performed during a study. The relation between heart transformation and patients' survival time was then investigated. As another example, in the AIDS Clinical Trial Group (ACTG) 181 dataset, patients were inspected at irregular intervals to determine is they had a cytomegalovirus (CMV) shedding occurred in either urine or blood. The main interest is to investigate the relation between shedding time and AIDS related disease.

Compared to the above two examples, a binary time-varying covariate in the former example was completely defined since a heart transplantation time was exactly known; however, the time of shedding occurrence in the latter example was not exactly observed because the occurrence was only detected by a lab examination or clinician's declaration. This kind of incomplete information makes it unable to apply a partial likelihood directly since uncertainty of a covariate affects the contribution of subjects in the risk set.

Let $Z_i(t)$ be a binary time-varying covariate of a subject i $(= 1, \ldots, n)$. Denote X_i as a changing time with a distribution function $G(x) = \Pr(X \leq x)$. Then, $Z_i(t) = 0$ for $t < X_i$ and $Z_i(t) = 1$ for $t \geq X_i$. We assume that this condition is permanent once a subject's status changes. If a subject's status

[1] Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Korea.
E-mail: yjin@sookmyung.ac.kr

Published 31 July 2016 / journal homepage: http://csam.or.kr

was periodically observed, X_i is not exactly observable. Instead, we observe (XL_i, XR_i) satisfying $XL_i < X_i < XR_i$ where $Z_i(t) = 0$ for $t \le XL_i$ and $Z_i(t) = 1$ for $t \ge XR_i$. Meanwhile, $Z_i(t)$ has uncertain value at $XL_i < t < XR_i$. Now, the available data is composed of

$$\{T_i, \delta_i, (XL_i, XR_i),\ i = 1, \ldots, n\},$$

where $T_i = \min(C_i, \tilde{T}_i)$ and $\delta_i = I(\tilde{T}_i < C_i)$. A right censoring time C_i is assumed to be independent of a failure time $\tilde{T}_i$ and a changing time X_i. Furthermore, interval censoring times (XL_i, XR_i) are assumed to be noninformative to T_i, C_i and X_i. Then, a proportional hazard model is implemented to investigate the effect of a binary time varying covariate on a failure time. Given $Z_i(t)$,

$$\lambda_i(t|Z_i(t)) = \lambda_0(t)\exp(\beta Z_i(t)),$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β is a regression coefficient, respectively. Define $Y_i(t) = I(T_i > t)$ as a risk indicator. Given a complete information about $Z_i(t)$, a partial estimating equation is derived as follows

$$U_n(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i(T_i) - \frac{S_1(\beta, T_i|Z(T_i))}{S_0(\beta, T_i|Z(T_i))} \right\} = 0, \tag{1.1}$$

where $S_j(\beta, t|Z(t)) = (1/n)\sum_{l=1}^{n} Y_l(t) Z_l^j(t) e^{\beta Z_l(t)}$ for $j = 0, 1$. However, when an information about $Z_i(t)$ is incomplete, (1.1) cannot be directly applied.

Goggins *et al.* (1999) considered an interval censored changing time and applied an EM algorithm. Their method is now reviewed as follows. For an equivalence set $\{Q_j = [l_j, r_j),\ j = 1, \ldots, I\}$ constructed from $\{XL_i, XR_i,\ i = 1, \ldots, n\}$ (Lindsey and Ryan, 1998), $g_j = G(l_j) - G(r_{j-1})$ denotes a corresponding probability mass function. In the E-step, M's sets of imputed values $j^r = (j_x^r(1), \ldots, j_x^r(n),\ r = 1, \ldots, M)$ are drawn from the conditional probability. Then, using such updated changing times, the corresponding $Z^{(j^r)}(t)$ are redefined. A joint likelihood is the product of a (conditional) partial likelihood of failure times and a marginal distribution of changing times as follows

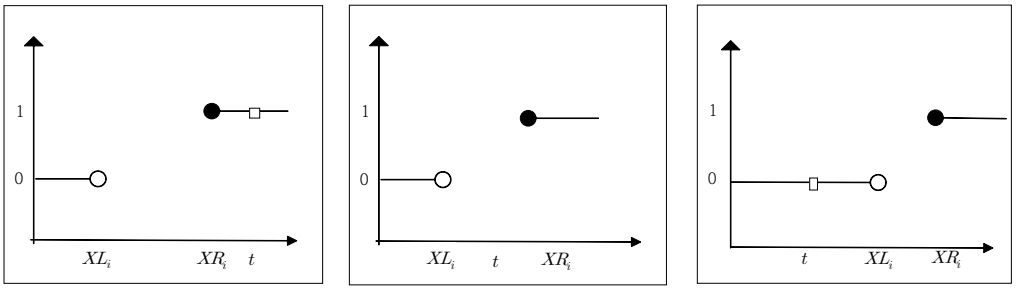
$$L(\beta, g|j) = \prod_{i=1}^{n} \left(\frac{e^{\beta z_i^{(j)}(t_i)}}{\sum_{k=1}^{n} Y_k(t_i) e^{\beta z_k^{(j)}(t_i)}} \right) \prod_{l=1}^{I} g_{j_x(l)}. \tag{1.2}$$

In the M-step, β and $g = \{g_1, \ldots, g_I\}$ are estimated by maximizing a partial likelihood and by employing sample proportions, respectively. However, the computation seems to be burdensome and a variance estimation is computationally difficult.

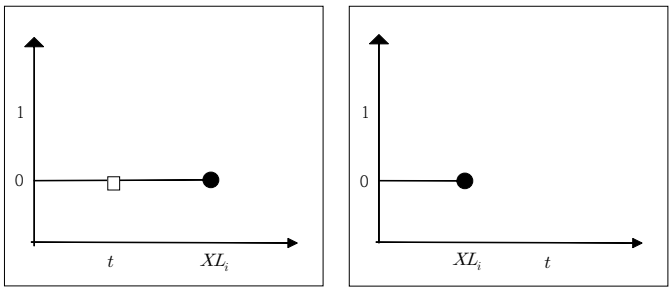
In this paper, we suggest a modified estimating equation by extending the method for a missing covariate problem. Section 2 presents a suggested method and Section 3 shows simulation results to evaluate the performance of the suggested method and a real data analysis appears in Section 4. Section 5 discusses the extension of the suggested method.

2. A modified estimating equation

In this study, we employ a modified estimating equation motivated by a method for a missing covariate. When a covariate is missing for some subjects, Zhou and Pepe (1995) redefined a relative risk function. Now, their study is briefly summarized as follows. Let $\tilde{\eta}$ be a missing indicator which has a value zero if $\tilde{Z}$ is missing and one otherwise. Consider an auxiliary covariate W which is



(a) $\eta_i(t) = 1,\ Z_i(t) = 1$ (b) $\eta_i(t) = 0,\ Z_i(t) = ?$ (c) $\eta_i(t) = 1,\ Z(t) = 0$



(d) $\eta_i(t) = 1,\ Z_i(t) = 0$ (e) $\eta_i(t) = 0,\ Z_i(t) = ?$

Figure 1: *Flow of interval censored time-varying binary covariate.*

completely observable and a categorical variable. For a subject i, a hazard function is defined as $\lambda_i(t;\tilde{Z},W) = \lambda_0(t)\exp(\beta_1\tilde{Z}_i + \beta_2 W_i) = \lambda_0(t)r_i(\beta,t)$, where $\beta = (\beta_1,\beta_2)$. Now, a relative risk function is defined as $R_i(\beta,t) = \tilde{\eta}_i r_i(\beta,t) + (1-\tilde{\eta}_i)\bar{r}_i(\beta,t)$, where $\bar{r}_i(\beta,t) = \exp(\beta_2 W_i)E(\exp(\beta_1\tilde{Z}_i)|T_i > t, W_i)$. Then, using an information of W_i, $\bar{r}_i(\beta,t)$ is estimated as follows

$$\hat{\bar{r}}_i(\beta,t) = \exp(\beta_2 W_i) \times \frac{\sum_{l=1}^{n} \tilde{\eta}_l I(T_l \le t, W_l = W_i)\exp(\beta_1 Z_l)}{\sum_{l=1}^{n} \tilde{\eta}_l I(T_l \le t, W_l = W_i)}.$$

Therefore, the contribution of a missing covariate to relative risk is replaced with a mean of relative risks calculated from the subjects with same values of W. Then the following estimated partial likelihood (EPL) score function is suggested

$$\tilde{U}_E(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \frac{\hat{R}_i^{(1)}(\beta,T_i)}{\hat{R}_i^{(0)}(\beta,T_i)} - \frac{\tilde{S}^{(1)}(\beta,T_i)}{\tilde{S}^{(0)}(\beta,T_i)} \right\} = 0, \tag{2.1}$$

where $\hat{R}_i^{(0)}(\beta,t) = \tilde{\eta}_i r_i(\beta,t) + (1-\tilde{\eta}_i)\hat{\bar{r}}_i(\beta,t)$, $\hat{R}_i^{(1)}(\beta,t) = \partial\hat{R}_i^{(0)}(\beta,t)/\partial\beta$, and $\tilde{S}^{(m)}(\beta,t) = \sum_{i=1}^{n} Y_i(t)\hat{R}_i^{(m)}(\beta,t)$ for $m = 0, 1$. Thus, W plays an important role to estimate relative risk. However, the above approach is inapplicable to our case since there is no auxiliary covariate in our dataset.

Moreover, $Z(t)$ depends on a changing time X which is either interval censored or right censored. Figure 1 shows $Z(t)$'s under five cases according to a censoring type and the relation between (XL, XR) and t. In detail, (a) $Z(t) = 1$ for $XL < XR < t$, (c) $Z(t) = 0$ for both $t < XL < XR$ and (d) $t < XL$, respectively. Meanwhile, $Z(t)$ has an uncertain value under (b) $XL < t < XR$ and (e) $XL < t$.

According to the certainty of $Z(t)$, $\eta(t)$ is defined. That is, cases (a), (c) and (d) giving either $Z(t) = 0$ or $Z(t) = 1$ result in $\eta(t) = 1$. Cases (b) and (e) with uncertain value of $Z(t)$ has $\eta(t) = 0$.

To solve such an uncertain status of $Z(t)$, a pseudo changing time is implemented. For finding suitable ones, take $d_k,\ k = 1, \ldots, I$, where $d_k = (q_k + p_k)/2$ is the midpoint of equivalence set $\{Q_k = (q_k, p_k),\ j = 1, \ldots, I\}$. Then a modified relative risk function is defined as

$$\tilde{R}_l^{(0)}(T,\beta) = \left\{ \eta_l(T) e^{\beta Z_l(T)} + (1 - \eta_l(T)) \frac{\sum_{k=1}^{I} \alpha_{lk} w_{lk} e^{\beta \tilde{Z}_{lk}(T)}}{\sum_{k=1}^{I} \alpha_{lk} w_{lk}} \right\}$$

and $\tilde{R}_l^{(1)}(T_i,\beta) = \partial \tilde{R}_l^{(0)}(T_i,\beta)/\partial\beta$ where a weight is defined as $w_{lk} = \Pr(X_l \in Q_k)$ and estimated with $\hat{w}_{lk} = \alpha_{lk}\hat{g}_k / \sum_{j=1}^{I} \alpha_{lj}\hat{g}_j$, where $\alpha_{lk} = I[Q_k \in (XL_l, XR_l)]$. Now, define a pseudo binary time-varying covariate as $\tilde{Z}_{lk}(t) = I(t \geq d_k)\alpha_{lk}$. Then $\tilde{Z}_{lk}(t)$ at risk set changes a value according to the relation between failure time t and an interval censored covariate (XL_l, XR_l).

A modified estimating equation is derived as follows

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \delta_i \left[\frac{\tilde{R}_i^{(1)}(T_i,\beta)}{\tilde{R}_i^{(0)}(T_i,\beta)} - \frac{\sum_{l=1}^{n} Y_l(T_i)\tilde{R}_l^{(1)}(T_i,\beta)}{\sum_{l=1}^{n} Y_l(T_i)\tilde{R}_l^{(0)}(T_i,\beta)} \right] = 0, \tag{2.2}$$

where to estimate g, a self-consistency algorithm (Turnbull, 1976) is applied. That is, $\{g_j, j = 1, \ldots, I\}$ is estimated by maximizing

$$L_s(g) = \prod_{i=1}^{n} [S(XL_i) - S(XR_i)] = \prod_{i=1}^{n} \sum_{k=1}^{I} \alpha_{ik} g_k.$$

Now, β is estimated by maximizing (2.2) and the variance is estimated with an inverse of a Hessian matrix.

3. Simulation

Austin (2012)'s method is applied to generate a time-varying binary covariate. Denote X_0 as a changing time at which the time-varying covariate changes a status. That is, $Z(t) = 0$ for $t < X_0$ and $Z(t) = 1$ for $t \geq X_0$. For generating X_0, an exponential distribution is applied, $X_0 \sim \exp(\lambda)$ and then a simulated survival time is defined as

$$T = \begin{cases} \dfrac{-\log(u)}{\lambda}, & \text{if } -\log(u) < \lambda X_0, \\ \dfrac{-\log(u) - \lambda X_0 + \lambda \exp(\beta) X_0}{\lambda \exp(\beta)}, & \text{if } -\log(u) \geq \lambda X_0, \end{cases}$$

where $u \sim U(0,1)$, $\beta = 0.3$ and $\lambda = 0.5$. In order to generate an interval censored covariate, the following procedure is applied. For each subject, a different number of inspection time is applied, $h_i \sim$ unif$(10, 15)$, where unif denotes a discrete uniform distribution. Then generate $w_l,\ l = 1, \ldots, h_i$ from $U(0, v)$ and sort these values. Now find two values satisfying $w_{(k-1)} < X_0 < w_{(k)},\ k = 1, \ldots, h_i$ and $w_{(0)} = 0$. Then set $XL = w_{(k-1)}$ and $XR = w_{(k)}$. In this simulation, we consider two settings depending on the length between XL and XR, (a wide interval and a narrow one) and these are controlled by v values. In order to evaluate the suggested method, we compare with the method using an exact changing time and two simpler procedures such as right point imputation and midpoint imputation

Table 1: Simulation result for wide intervals

Censoring rate		$n = 100$				$n = 200$			
		Exact	Midpoint	Right point	Suggested	Exact	Midpoint	Right point	Suggested
30%	Bias	0.013	0.113	0.144	0.030	0.003	0.085	0.121	0.050
	SE	0.270	0.315	0.372	0.323	0.190	0.220	0.259	0.225
	ESE	0.273	0.317	0.391	0.298	0.204	0.210	0.257	0.221
	95%CP	0.956	0.933	0.933	0.966	0.926	0.946	0.946	0.946
45%	Bias	0. 018	0.116	0.128	0.028	0.001	0.086	0.105	0.072
	SE	0.296	0.365	0.447	0.361	0.205	0.250	0.306	0.251
	ESE	0.289	0.364	0.439	0.330	0.194	0.241	0.313	0.220
	95%CP	0.963	0.940	0.870	0.963	0.963	0.950	0.943	0.950

SE = standard errors; ESE = empirical standard errors; 95%CP = 95% coverage probabilities.

Table 2: Simulation result for narrow intervals

Censoring rate		$n = 100$				$n = 200$			
		Exact	Midpoint	Right point	Suggested	Exact	Midpoint	Right point	Suggested
30%	Bias	0.003	0.068	0.104	0.001	0.002	0.065	0.087	0.009
	SE	0.268	0.299	0.341	0.308	0.188	0.109	0.237	0.214
	ESE	0.266	0.304	0.356	0.300	0.185	0.228	0.252	0.205
	95%CP	0.946	0.946	0.930	0.960	0.953	0.920	0.923	0.966
45%	Bias	0.011	0.087	0.111	0.017	0.012	0.094	0.100	0.001
	SE	0.296	0.342	0.411	0.347	0.200	0.240	0.279	0.240
	ESE	0.316	0.337	0.455	0.336	0.215	0.250	0.309	0.243
	95%CP	0.947	0.963	0.937	0.963	0.943	0.963	0.916	0.960

SE = standard errors; ESE = empirical standard errors; 95%CP = 95% coverage probabilities.

where interval censored covariate is replaced by the right points and the midpoint, respectively. Two sample sizes ($n = 100, 200$) are considered and two different censoring rates ($p = 0.3, 0.45$) are implemented. Table 1 summarizes the simulation results from 300 replicates when interval (XL, XR) is wide. It shows biases, means of standard errors (SE), empirical standard errors (ESE), and 95% coverage probabilities (95%CP) of four methods.

Compared with two imputation methods, the suggested method provides smaller biases and smaller standard errors. However, all empirical coverage probabilities are close to the nominal levels. However, midpoint methods have better coverage probabilities that the suggested method at wide intervals where the standard errors of the suggested method are overestimated. Now, in order to check the effect of such imputed values, the relation among a true changing time X_0, a failure time T and imputed value X_M is investigated. For example, $X_0 < T < X_M$ results in $Z(T) = 0$ with an imputed value X_M even though a true value of $Z(T)$ is one. Such mismatch probabilities are 12% and 24% at midpoint and right point imputation, respectively. These proportions explain why a right point imputation has larger bias than a midpoint imputation. Table 2 shows the simulation result under a narrow interval of (XL, XR). Mismatch probabilities are again 8% and 20%, respectively which are smaller than those at a wide interval. By comparing with the results of a wide interval, biases and standard errors decrease for all methods. Furthermore, the standard errors decrease as sample size increases and the decreasing rate is approximately $\sqrt{n}$.

4. Data analysis

Goggins *et al.* (1999) investigated the relation between CMV shedding in urine and blood and the onset of active CMV end-organ disease. We also analyze this relation from 193 patients' data. While the disease onset times are known, the CMV shedding times are either right censored, interval censored

Table 3: Hazard of CMV shedding on CMV disease onset time

	Right	Midpoint	Suggested
$\hat{\beta}$	0.805	0.572	1.235
$\hat{\mathrm{SE}}(\hat{\beta})$	0.184	0.180	0.218
z	4.370	3.170	5.661
p-value	< 0.001	0.002	< 0.001

CMV = cytomegalovirus; SE = standard errors.

or left censored since CMV shedding was detected by screening tests at clinic visits. Patients have different numbers of visiting and their lag times were also irregular. For example, left censored shedding occurred when patients already experienced shedding at the time they entered the study. Also, shedding time is regarded as a right censored at a failure time when failure occurs without shedding. Among the 193 patients, 39 had no CMV shedding until CMV disease onset time. Table 3 shows the results of three methods. A significant and positive relation between CMV shedding time and the onset of active CMV disease is found in the three methods and these results are consistent with Goggins *et al.* (1999)'s data analysis. However, the results with the midpoint and right point imputation have smaller $\hat{\beta}$ values that can be explained by a smaller portion of $Z = 1$ as the simulation results where the estimates based on the imputation are negative-biased. This result is denoted as an attenuation by Goggins *et al.* (1999).

5. Discussion

We consider a binary time-varying covariate in a context of a failure time regression model. In particular, our interest is to estimate a regression coefficient when a changing time is incompletely observed. We proposed a modified estimating equation. Unlike other approaches to replace unknown changing time with a value, several pseudo values are imputed with corresponding weights into a partial score function. Simulation studies show that the proposed method performs better than simpler imputation methods which provide under-estimated ones. Furthermore, the suggested method could be easily implemented with statistical packages once weights are calculated.

As another type of interval censored covariate can appear as a continuous covariate. Gómez *et al.* (2003) considered an exponential family regression model when a continuous covariate is interval censored. As the extension of this approach, a statistical model of interval censored covariate at a right censored failure time was proposed. A parametric survival distribution was assumed and a corresponding likelihood was derived by Langohr *et al.* (2004). They constructed a joint likelihood of both failure time and an interval censored continuous covariate. The likelihood was reexpressed as weighting form under the assumption of a discrete support for a covariate distribution. Zhao *et al.* (2004) also considered an estimating equation for a case where both response failure time and covariate are interval censored. Their approach was based on the estimating equation approach which Sun *et al.* (1999) suggested to analyze doubly interval censored data. In a context of recurrent event data, Chen and Cook (2005) applied a marker process to time dependent interval censored data and proposed a joint analysis of marker process and recurrent events time process. Alternatively, a multi-state model regarding a binary time-varying covariate as a state has also been studied in several examples (Commenges, 2002).

Acknowledgements

The author thanks Dr. Dianne Finkelstein for providing the dataset. This work was supported by the National Research Foundation of Korea (NRF) grant (NRF-2014R1A2A2A01003567).

References

Austin PC (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates, *Statistics in Medicine*, **31**, 3946–3958.

Chen EB and Cook RJ (2005). Regression modeling with recurrent event and time-dependent interval censored marker data, *Lifetime Data Analysis*, **9**, 275–291.

Commenges D (2002). Inference for multi-state models from interval-censored data, *Statistical Methods in Medical Research*, **11**, 167–182.

Goggins WB, Finkelstein DM, and Zaslavsky AM (1999). Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval-censored, *Biometrics*, **55**, 445–451.

Gómez G, Espinal A, and Lagakos SW (2003). Inference for a linear regression model with an interval-censored covariate, *Statistics in Medicine*, **22**, 409–425.

Langohr K, Gómez G, and Muga R (2004). A parametric survival model with an interval-censored covariate, *Statistics in Medicine*, **23**, 3159–3175.

Lindsey JC and Ryan LM (1998). Methods for interval censored data, *Statistics in Medicine*, **17**, 219–238.

Sun J (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

Sun J, Liao Q, and Pagano M (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies, *Biometrics*, **55**, 909–914.

Turnbull BW (1976). The empirical distribution function with arbitrarily grouped censored and truncated data, *Journal of Royal Statistical Society B*, **38**, 290–295.

Zhao X, Lim HJ, and Sun J (2004). Estimating equation approach for regression analysis of failure time data in the presence of interval-censoring, *Journal of Statistical Planning and Inference*, **129**, 145–157.

Zhou H and Pepe MS (1995). Auxiliary covariate data in failure time regression analysis, *Biometrika*, **82**, 139–149.

Received June 5, 2016; Revised July 17, 2016; Accepted July 17, 2016