Statistics and Its Interface Volume 11 (2018) 91–97

Regression analysis of incomplete data from event history studies with the proportional rates model

Guanglei Yu, Liang Zhu, Jianguo Sun*, and Leslie L. Robison

Event history studies occur in many fields including epidemiology, sociology, and medical studies. They focus on the occurrences of some events of interest on subjects over time. One special type of data arising from such studies is incomplete mixed data, which is the mixed recurrent event data and panel count data. To deal with such type of data, we propose a proportional rates model and present a multiple imputation-based estimation procedure. One advantage of the proposed marginal model approach is that it can be easily implemented. To assess the performance of the procedure, a simulation study is conducted and indicates that it performs well for practical situations and can be more efficient than the existing method. The methodology is applied to a set of mixed data from a longitudinal cohort study.

Keywords and phrases: Incomplete data, Marginal model, Multiple imputation, Proportional rates model.

1. INTRODUCTION

This paper discusses regression analysis of a type of incomplete data arising from event history studies, commonly conducted in demography, economics, medical studies and social sciences. More specifically, we consider mixed recurrent event and panel count data or the mixture of recurrent event data [1] and panel count data [2]. The former is a type of complete data and usually means that all study subjects are observed continuously over the whole study period, while the latter arises if each study subject is observed only at discrete points. In the latter case, only the number of occurrences of events of interest between consecutive observation time points are known or recorded. By mixed data, we mean that each study subject may be observed continuously during the whole study period, continuously over some study periods and at some time points otherwise, or only at some discrete time points. The observed data consist both complete and incomplete sample paths. Note that these time periods may be different for different subjects. If each subject gives either complete or incomplete sample paths over the whole study period, the observed data are commonly referred to as type I mixed data. Otherwise, they are usually referred to as type II mixed data [3]. It is apparent that the latter is much more complicated than the former.

*Corresponding author.

A great deal of literature has been established for the analysis of both recurrent event data and panel count data separately. In particular, [1] and [2] provided relatively complete reviews and references of the two fields, respectively. In contrast, there exist only a couple of methods that apply to the analysis of mixed recurrent event and panel count data. One major problem with respect to the analysis of mixed data is combining the two different types of data structures together. The first paper to discuss statistical analysis of such data was given by [3], which proposed some estimating equation-based methods under the proportional means model. In particular, they presented a set of mixed recurrent event and panel count data arising from a multi-center longitudinal cohort study, the Childhood Cancer Survivor Study (CCSS) (http://ccss.stjude.org; [4]). One of the major goals of the study is to assess the long-term effects, if any, of childhood cancer and cancer treatments on the subsequent reproductive functions of childhood cancer survivors (more details will be provided below). Following [3], [5] considered the same problem but only for type I mixed data, while [6] investigated regression analysis of mixed recurrent event and panel count data arising from the additive rates model.

Multiple imputation is a commonly used approach when there exists missing or incomplete data. The idea behind it is to fill or replace the missing or unobserved values by some imputed values [7]. It is well-known that one key advantage of the method is that it can be easily implemented and allows one to employ the existing methods for the analysis of the corresponding complete data. Also it has been studied by many authors theoretically and applied to many areas including failure time data analysis [7, 8, 9, 10, 11, 12, 13, 14]. For example, [14] developed two imputation procedures for regression analysis of right-censored failure time data under the accelerated failure time model, and [11] discussed the use of the multiple imputation approach for regression analysis of interval-censored failure time data arising from the proportional hazards model. In the following, we will treat the panel count data as missing data and develop a multiple imputation procedure that imputes the unobserved occurrence times of the recurrent events of interest. The procedure is essentially a marginal model approach.

The remainder of this paper is organized as follows. In Section 2, we first introduce the notation and the proportional rates model and then briefly review the inference pro-

cedures proposed by [15] and [16] for the analysis of recurrent event data. Section 3 presents the developed multiple imputation procedure for regression analysis of mixed recurrent event and panel count data and in the method, the idea discussed in [11] is used. The procedure applies to both missing completely at random (MCAR) and missing at random (MAR) mechanisms. In Section 4, we report some results obtained from a simulation study conducted to assess the performance of the multiple imputation procedure. They indicate that the proposed approach performs well for practical situations and can be more efficient than that given in [3]. The method is applied to the CCSS mentioned above in Section 5 and Section 6 contains some discussion and concluding remarks.

2. ESTIMATION PROCEDURE FOR RECURRENT EVENT DATA

Consider an event history study that consists of n independent subjects. For subject i, let $N_i^*(t)$ denote the total number of events of interest that have occurred up to time t, $0 \leqslant t \leqslant \tau$, where τ denotes the study length, and suppose that there exists a vector of covariates denoted by $\mathbf{X}_i$. Also suppose that for each subject there exists an independent follow-up time C_i^*, and define $C_i = (C_i^* \wedge \tau)$, $N_i(t) = N_i^*(t \wedge C_i)$, and $Y_i(t) = I(t \leqslant C_i)$. To describe the covariate effect on $N_i^*(t)$, we will assume that given $\mathbf{X}_i$, $N_i^*(t)$ follows the proportional rates model

$$E\{dN_i^*(t)|\boldsymbol{X}_i\} = \lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i)dt\,, \tag{1}$$

where $\lambda_0(t)$ denotes an unspecified baseline rate function and $\boldsymbol{\beta}$ a vector of regression parameters. Define the baseline mean function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. In the following, we assume that the main goal is to estimate $\boldsymbol{\beta}$.

For estimation of regression parameters $\boldsymbol{\beta}$, we will assume that one observes recurrent event data on the $N_i(t)$'s. That is, all occurrence times of the recurrent events of interest within the follow-up periods are known. In this case, [15] suggested to use the estimator $\hat{\boldsymbol{\beta}}_r$ defined as the solution to the estimating equation

$$U_n(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \left\{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t;\boldsymbol{\beta})\right\} dN_i(t) = 0\,,$$

where $\bar{\boldsymbol{X}}(t;\boldsymbol{\beta}) = S^{(1)}(t;\boldsymbol{\beta})/S^{(0)}(t;\boldsymbol{\beta})$ with

$$S^{(j)}(t;\boldsymbol{\beta}) = \sum_{i=1}^n Y_i(t)\,\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i)\,\boldsymbol{X}_i^{\otimes j}\,,$$

$j = 0, 1, 2$, and $\boldsymbol{a}^{\otimes 0} = 1, \boldsymbol{a}^{\otimes 1} = \boldsymbol{a}, \boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ for a vector $\boldsymbol{a}$. Furthermore, [16] showed that the distribution of $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_0)$ can be asymptotically approximated by the multivariate normal distribution with mean zero and the covariance matrix estimated by $\hat{\boldsymbol{\Sigma}}_r = \hat{\boldsymbol{A}}_r^{-1}\hat{\boldsymbol{\Gamma}}_r\hat{\boldsymbol{A}}_r^{-1}$. Here $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{A}}_r = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t;\hat{\boldsymbol{\beta}}_r)\}^{\otimes 2} Y_i(t) \times \exp\{\hat{\boldsymbol{\beta}}_r^T\boldsymbol{X}_i(t)\}d\hat{\Lambda}_{0r}(t)\,,$$

and

$$\hat{\boldsymbol{\Gamma}}_r = \frac{1}{n}\sum_{i=1}^n [\int_0^\tau \{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t;\hat{\boldsymbol{\beta}}_r)\}d\hat{M}_i(t)]^{\otimes 2}\,,$$

where $\hat{\Lambda}_{0r}(t)$ and $\hat{M}_i(t)$ are given as:

$$\hat{\Lambda}_{0r}(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{S^{(0)}(s;\hat{\boldsymbol{\beta}}_r)}\,,$$

and

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)\exp\{\hat{\boldsymbol{\beta}}_r^T\boldsymbol{X}_i\}d\hat{\Lambda}_{0r}(s)\,.$$

Note that as an alternative from [3], one can replace $\hat{\boldsymbol{A}}_r$ above by the following estimator

$$\hat{\boldsymbol{A}}_r = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t;\hat{\boldsymbol{\beta}}_r)\}^{\otimes 2} dN_i(t)\,.$$

In the next section, we will generalize the estimation procedure above to cases of mixed recurrent event and panel count data by using the multiple imputation approach.

3. ESTIMATION PROCEDURE FOR MIXED RECURRENT EVENT AND PANEL COUNT DATA

Now we suppose that only mixed recurrent event and panel count data are available and for estimation of regression parameters $\boldsymbol{\beta}$ in model (1), we will present a multiple imputation-based estimation procedure. For this, suppose that for each subject there exists a sequence of time points $0 = T_{i0} < T_{i1} < \cdots < T_{iK_i}$ such that within $(T_{i,j-1}, T_{ij}]$, either subject i is continuously observed, and thus gives complete data, or only the number of occurrences of the events from subject i is known, $j = 1, ..., K_i$, $i = 1, ..., n$. Also suppose that one observes an indicator function $r_i(t)$ with $r_i(t) = 1$ for $t \in (T_{i,j-1}, T_{ij}]$, if complete data are available from subject i over the interval, and $r_i(t) = 0$ otherwise. One can see that the indicator functions $r_i(t)$'s serve as missing indicators and in the following, we assume that the missing is either MCAR or MAR. Furthermore, it is easy to see that if the $r_i(t)$'s are independent of time t, then we have type I mixed data and otherwise, one observes type II data. Define $O_i^*(t) = \sum_{j=1}^{K_i} I(T_{ij} \leq t)$, $i = 1, ..., n$. In the following, we will assume that $N_i^*(t)$, $O_i^*(t)$, C_i^* and $r_i(t)$ are mutually independent, conditional on $\boldsymbol{X}_i$.

To develop the multiple imputation estimation procedures, the key is to impute the unobserved occurrence times of the recurrent events of interest over all inter-

vals $(T_{i,j-1}, T_{ij}]$, $j = 1, ..., K_i$, $i = 1, ..., n$, within which only incomplete data are available. For such intervals, let N_{ij} denote the number of occurrences of the events over $(T_{i,j-1}, T_{ij}]$ from subject i. For the generation of the imputed occurrence times, note that if $N_i^*(t)$ is a Poisson process, following [17], we can show that given $\{ (T_{i,j-1}, T_{ij}], N_{ij}, \boldsymbol{X}_i \}$, the occurrence times of the N_{ij} recurrent events are the order statistics of i.i.d random variables from the density function

$$f_{ij}(t) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)\lambda_0(t) I(T_{i,j-1} < t \leqslant T_{ij})}{\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)(\Lambda_0(T_{ij}) - \Lambda_0(T_{i,j-1}))} = \frac{\lambda_0(t) I(T_{i,j-1} < t \leqslant T_{ij})}{(\Lambda_0(T_{ij}) - \Lambda_0(T_{i,j-1}))} .$$

By following [11] and [14], this motivates the following imputation procedure.

Let B be a chosen integer for the number of imputed data sets and $\hat{\boldsymbol{\beta}}^{(l)}$ and $\hat{\Lambda}_0^{(l)}(t)$ denote the estimators of $\boldsymbol{\beta}$ and $\Lambda_0(t)$, respectively, obtained in the lth iteration of the iterative procedure below.

- **Step 1:** Choose initial estimators $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\Lambda}_0^{(0)}(t)$ of $\boldsymbol{\beta}$ and $\Lambda_0(t)$, respectively.
- **Step 2:** At the lth iteration and for each $b = 1, ..., B$, define the bth set of imputed recurrent event data as follows. For any $i = 1, ..., n$ and $j = 1, ..., K_i$, if $r_i(t) = 0$ for $t \in (T_{i,j-1}, T_{ij}]$, define the N_{ij} occurrence times of the event as the order statistics of a random sample of size N_{ij} drawn from the candidate time points $\{ s_{ij,1}, ..., s_{ij,m_{ij}} \}$ within the interval $(T_{i,j-1}, T_{ij}]$ with the probability mass $\{ p_{ij,1}, ..., p_{ij,m_{ij}} \}$, where the candidate time points are the combined distinct time points of both the observed recurrent event times of the subjects other than subject i, and the imputed event time points falling within this interval among B data sets from the $(l-1)$th iteration, and

$$p_{ij,q} = \frac{d\hat{\Lambda}_0^{(l-1)}(s_{ij,q})}{\sum_{r=1}^{m_{ij}} d\hat{\Lambda}_0^{(l-1)}(s_{ij,r})} , \; q = 1, ..., m_{ij} .$$

- **Step 3:** For each $b = 1, ..., B$, define the estimators $\hat{\boldsymbol{\beta}}_b^{(l)}$, $\hat{\Lambda}_{0b}^{(l)}(t)$ and $\hat{\boldsymbol{\Sigma}}_b^{(l)}$ as the estimators $\hat{\boldsymbol{\beta}}_r$, $\hat{\Lambda}_{0r}(t)$ and $\frac{1}{n}\hat{\boldsymbol{\Sigma}}_r$ defined in the previous section based on the bth imputed recurrent event data defined in the previous step.
- **Step 4:** Obtain the updated estimators of $\boldsymbol{\beta}$ and $\Lambda_0(t)$ by:

$$\hat{\boldsymbol{\beta}}^{(l)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_b^{(l)} , \; \hat{\Lambda}_0^{(l)}(t) = \frac{1}{B}\sum_{b=1}^{B} \hat{\Lambda}_{0b}^{(l)}(t) ,$$

and the estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}^{(l)}$ by:

$$\hat{\boldsymbol{\Sigma}}^{(l)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\Sigma}}_b^{(l)} + \left(1 + \frac{1}{B}\right) \sum_{b=1}^{B} \frac{(\hat{\boldsymbol{\beta}}_b^{(l)} - \hat{\boldsymbol{\beta}}^{(l)})(\hat{\boldsymbol{\beta}}_b^{(l)} - \hat{\boldsymbol{\beta}}^{(l)})'}{B-1} .$$

- **Step 5:** Return to Step 2 until the convergence.

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(t)$ denote the estimators of $\boldsymbol{\beta}$ and $\Lambda_0(t)$ obtained above and $\hat{\boldsymbol{\Sigma}}$ the resulting covariance estimator for $\hat{\boldsymbol{\beta}}$. Then under some mild regularity conditions, it is expected that $\hat{\boldsymbol{\beta}}$ is consistent and the distribution of $\hat{\boldsymbol{\beta}}$ can be asymptotically approximated by the normal distribution with mean $\boldsymbol{\beta}_0$ and a variance-covariance matrix that can be estimated by $\hat{\boldsymbol{\Sigma}}$. To implement the iterative algorithm above, two issues need to be discussed. They are the selection of initial estimators and the convergence. For the former, in any interval $(T_{i,j-1}, T_{ij}]$ with $r_i(t) = 0$, we first generated the unobserved event times from the uniform distribution over $(T_{i,j-1}, T_{ij}]$ and then set $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\Lambda}_0^{(0)}(t)$ as average of $\hat{\boldsymbol{\beta}}_r$ and $\hat{\Lambda}_{0r}(t)$ defined in the previous section based on the multiple generated recurrent event data sets. With respect to the convergence, it is apparent that one can compare the estimators of both $\boldsymbol{\beta}$ and $\Lambda_0(t)$ from two consecutive iterative steps and stop the iteration if the overall difference is small enough. Alternatively, if one is only interested in estimation of $\boldsymbol{\beta}$, we may only compare the estimators $\hat{\boldsymbol{\beta}}^{(l-1)}$ and $\hat{\boldsymbol{\beta}}^{(l)}$ and stop the iteration if $||\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}|| \leqslant \epsilon$, where ϵ is a prespecified positive constant. In the numerical study below, the convergence did not seem to be an issue.

4. A SIMULATION STUDY

A simulation study was conducted to assess the performance of the estimation procedure proposed in the previous section, and in the study we considered both types I and II of mixed recurrent event and panel count data. To generate the simulated data, following [3], we assumed below that there was one covariate X_i following the Bernoulli distribution with the success probability 0.5 and generated the follow-up time C_i^* from the uniform distribution over $(\tau/2, \tau)$ with $\tau = 1$. For the underlying recurrent event process $N_i^*(t)$, we considered two situations. One is the Poisson process with the mean function $\Lambda_0(t)\exp(\beta^T X_i)$ and the other is the mixed Poisson process with the mean function $\nu_i \Lambda_0(t)\exp(\beta^T X_i)$, where $\Lambda_0(t) = 3t$ and ν_i is a latent variable following the Gamma distribution with mean 1 and variance 0.25. To generate mixed recurrent event and panel count data, we first generated a sequence of time points as the T_{ij}'s from the Poisson process with mean function $\mu_0(t) = 3t$. For type I mixed data, note that each subject is observed either continuously or only at discrete times during the whole follow-up study. Therefore, the type indicator functions $r_i(t)$'s over all intervals within each subject are the same and were generated from the Bernoulli distribution with the success probability p_r.

For the generation of type II mixed recurrent event and panel count data, we first generated a sequence of time

points the same way as for type I mixed data, and then within each interval the type indicator $r_i(t)$ was generated from the Bernoulli distribution with success probability p_r so that $r_i(t)$ can be different over different time periods for each subject. With respect to p_r, for both types of mixed data, we also considered two situations corresponding to two different missing mechanisms. One is to take $p_r = 0.2$, 0.5 or 0.8, the same constants for all subjects, and the other is to let $p_r = 0.1$, 0.4 or 0.7 for the subjects with $X_i = 0$ and 0.3, 0.6 or 0.9 otherwise. That is, p_r depends on the covariate. The results below are based on $n = 100$ and $B = 10$ in the imputation procedure with 1000 replications.

Tables 1 and 2 present the results on estimation of β with $\beta_0 = -0.5$, 0 or 0.5 and p_r being the same for all subjects. Table 1 is the type I mixed recurrent event and panel count data and Table 2 is the type II mixed data. They include the estimated bias (BIAS) given by the average of the estimates minus the true value, the average of the estimated standard errors (ESE), the sample standard error of the estimators (SSE), and the 95% empirical coverage probabilities (CP). For comparison, we also applied the estimation procedure given in [3] to the simulated data and included the estimation results in the tables. One can see from the tables that the results suggest that the proposed estimator seems to be unbiased and the variance estimation appears to be reasonable. Also the normal approximation to the distribution of $\hat{\beta}$ seems to be appropriate. Furthermore, they indicate that the proposed estimator is more efficient and performs better than that given in [3].

The results for estimation of β in the situation that p_r depends on the covariates are presented in Tables 3 and 4 with the other set-ups being the same as in Tables 1 and 2. As above, Table 3 is for type I mixed recurrent event and panel count data, whereas Table 4 is for type II mixed data. It is easy to see that they gave similar conclusions as above and again suggest that the proposed multiple imputation estimation procedure seems to be more efficient than that given in [3]. In addition, as in Tables 1 and 2, the results indicate that the proposed estimation procedure seems to be much less dependent on p_r than the method proposed by [3]. A possible reason for this may be due to the nature of the multiple imputation method, which essentially creates and makes inferences based on complete or recurrent event data, while [3] bases inferences on the incomplete or mixed data. To assess the normal approximation to the distribution of $\hat{\beta}$, we also obtained the quantile plots of the standardized $\hat{\beta}$ against the standard normal distribution, and they (not shown here) again indicate that the approximation seems to be appropriate for the situations considered here. We also investigated some other set-ups including the cases with continuous covariates or different functions for $\Lambda_0(t)$ and obtained similar simulation results. In particular, we performed some simulation studies to assess the possible effect of the number of imputations B on the performance of the proposed method and the results suggested that they were not sensitive to the selection of B.

5. AN APPLICATION

In this section, we apply the estimation procedure proposed in the previous sections to the CCSS described above. The study consists of childhood cancer survivors who were diagnosed between 1970 and 1986 and had survived more than 5 years since diagnosis, along with a random sample of their siblings serving as a control group. The study subjects were distributed a baseline summary questionnaire starting in 1996 about their pregnancy information such as the age range of the pregnancies as well as other related information. If a pregnancy was reported in the summary questionnaire, that participant received an additional detailed pregnancy questionnaire for each reported pregnancy for additional information such as the exact age of pregnancy. If a subject only returned the summary questionnaire, then we would have incomplete panel count data on her pregnancy process, while if the subject returned both questionnaires, complete recurrent event data would be available. Actually this is the case between 1996 and 2000 and in other words, we have type I mixed recurrent event and panel count data up to 2000. Overall up to 2007, there existed some subjects who returned the detailed pregnancy questionnaire over some periods, but not over other periods although had pregnancies. That is, overall we have type II mixed recurrent event and panel count data on the pregnancy process. As mentioned before, one of the CCSS objectives is to determine the long-term effects, if any, of childhood cancer and its treatments on the pregnancy process or pregnancy outcomes.

In the analysis below, following [3], we will focus on the subgroup of 3966 female participants who were at least 25 years old in 1996, with 2765 being childhood cancer survivors and the others being their siblings. Define $X_i = 1$ if the ith subject is a survivor and $X_i = 0$ otherwise. First we considered the type I mixed data collected up to 2000. For these individuals, we have the averages of pregnancy counts being 1.498 and 2.049 for the survivor and sibling groups, respectively. The application of the proposed estimation procedure gave $\hat{\beta} = -0.319$ with the estimated standard error being 0.039, yielding a p-value of less than 0.0001 for testing no pregnancy process difference between the survivor and sibling groups. This suggests that the cancer survivors had significantly lower pregnancy rates than their siblings. In contrast, [3] gave the corresponding estimate of -0.128 with the estimated standard error of 0.034. Note that although the conclusions are similar, the estimated effect by the proposed method seems to be more significant.

Now we apply the proposed estimation procedure to the whole type II mixed data collected between 1996 and 2007. For them, the average pregnancy counts are 1.684 and 2.403 for the cancer survivors and their siblings, respectively, which are higher than the type I mixed data, as expected. By applying the proposed method, we obtained $\hat{\beta} = -0.324$ with the estimated standard error of 0.029. In contrast, [3] gave $\hat{\beta} = -0.247$ with the estimated standard error being 0.032. Again both approaches gave similar conclusions and indicated that the occurrence of childhood cancer and its

Table 1. Estimation of β based on type I mixed data with p_r independent on the covariate

			Proposed procedure				Zhu et al. (2013)			
$N_i^*(t)$	p_r	β_0	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
Poisson	0.2	0.5	0.009	0.145	0.149	0.943	0.011	0.221	0.235	0.931
		0	-0.010	0.161	0.165	0.948	-0.003	0.236	0.251	0.933
		-0.5	-0.011	0.187	0.184	0.952	-0.016	0.260	0.273	0.939
	0.5	0.5	0.003	0.145	0.150	0.948	0.007	0.209	0.227	0.924
		0	-0.002	0.161	0.158	0.950	0.007	0.224	0.236	0.925
		-0.5	-0.006	0.186	0.189	0.950	0	0.248	0.255	0.947
	0.8	0.5	0.006	0.144	0.150	0.941	0.002	0.176	0.192	0.916
		0	-0.011	0.161	0.170	0.931	-0.009	0.192	0.216	0.913
		-0.5	-0.002	0.187	0.192	0.939	-0.006	0.216	0.236	0.924
Mixed	0.2	0.5	-0.004	0.179	0.188	0.945	-0.007	0.250	0.265	0.939
Poisson		0	0	0.193	0.198	0.940	0.002	0.265	0.276	0.943
		-0.5	-0.003	0.213	0.223	0.936	-0.002	0.283	0.297	0.937
	0.5	0.5	0	0.179	0.189	0.935	0	0.239	0.264	0.925
		0	0.011	0.193	0.194	0.946	0.007	0.252	0.276	0.920
		-0.5	-0.010	0.213	0.226	0.937	-0.008	0.274	0.297	0.934
	0.8	0.5	0.010	0.179	0.184	0.945	0.018	0.209	0.217	0.927
		0	0.007	0.193	0.199	0.941	0.009	0.221	0.240	0.936
		-0.5	-0.006	0.215	0.228	0.936	-0.011	0.242	0.269	0.917

Table 2. Estimation of β based on type II mixed data with p_r independent on the covariate

			Proposed procedure				Zhu et al. (2013)			
$N_i^*(t)$	p_r	β_0	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
Poisson	0.2	0.5	0.003	0.145	0.146	0.951	0.002	0.218	0.233	0.932
		0	0	0.162	0.158	0.952	-0.002	0.231	0.238	0.940
		-0.5	-0.004	0.186	0.198	0.941	-0.001	0.254	0.275	0.920
	0.5	0.5	0.008	0.145	0.147	0.954	0.008	0.206	0.216	0.940
		0	-0.002	0.161	0.162	0.948	0.008	0.220	0.237	0.923
		-0.5	-0.001	0.185	0.189	0.943	0.007	0.242	0.245	0.949
	0.8	0.5	0.006	0.145	0.149	0.945	0.007	0.175	0.185	0.928
		0	-0.004	0.162	0.170	0.933	-0.006	0.191	0.209	0.928
		-0.5	-0.014	0.187	0.196	0.945	-0.008	0.214	0.231	0.927
Mixed	0.2	0.5	0.009	0.181	0.188	0.941	0.010	0.246	0.256	0.937
Poisson		0	-0.008	0.193	0.199	0.942	-0.012	0.259	0.267	0.934
		-0.5	-0.008	0.215	0.231	0.940	-0.023	0.281	0.306	0.926
	0.5	0.5	0.002	0.181	0.185	0.944	0.001	0.235	0.245	0.939
		0	0.009	0.194	0.198	0.942	0.008	0.248	0.264	0.942
		-0.5	-0.010	0.214	0.219	0.942	0.001	0.265	0.285	0.934
	0.8	0.5	0.003	0.179	0.188	0.931	0.007	0.205	0.224	0.925
		0	-0.004	0.194	0.196	0.947	0.004	0.219	0.235	0.933
		-0.5	-0.002	0.214	0.232	0.930	-0.001	0.238	0.255	0.930

treatments seemed to have significant effects in decreasing the pregnancy rate. Also as seen in the simulation study, the proposed method yielded more significant effects than that given in [3]. It is worth noting that in contrast to the proportional rates model discussed here, [3] considered the proportional means model and with time-invariant covariates, and the two models are equivalent.

6. CONCLUDING REMARKS

In this paper, we discussed regression analysis of mixed recurrent event and panel count data, a type of incomplete or missing data arising from event history studies. As described above, such data may occur in two forms, types I and II, and the former is a special case of the latter. For the analysis, we proposed a multiple imputation estimation procedure that converts the mixed data structure or incomplete data to complete recurrent event data by imputing the unobserved occurrence times. Note that the method is a marginal approach. It does not need joint modeling, and is valid under both MCAR and MAR missing mechanisms. Also note that although the idea for the imputation is borrowed from the Poisson process assumption, a working assumption, the

Table 3. Estimation of β based on type I mixed data with p_r dependent on the covariate

			Proposed procedure				Zhu et al. (2013)			
$N_i^*(t)$	p_r	β_0	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
Poisson	0.2	0.5	0.006	0.144	0.151	0.941	-0.002	0.228	0.243	0.929
		0	-0.001	0.162	0.164	0.953	-0.007	0.245	0.248	0.948
		-0.5	0.015	0.185	0.194	0.938	0.009	0.267	0.287	0.929
	0.5	0.5	0.003	0.144	0.149	0.947	0.005	0.215	0.226	0.942
		0	0.006	0.162	0.169	0.935	0.014	0.227	0.251	0.913
		-0.5	-0.006	0.186	0.190	0.934	-0.009	0.250	0.261	0.941
	0.8	0.5	0.003	0.145	0.146	0.943	0.010	0.180	0.205	0.908
		0	0.006	0.162	0.161	0.947	0.010	0.193	0.204	0.933
		-0.5	-0.004	0.186	0.189	0.947	-0.002	0.213	0.227	0.934
Mixed	0.2	0.5	0.003	0.179	0.183	0.945	-0.001	0.257	0.274	0.929
Poisson		0	0.009	0.193	0.196	0.947	0.012	0.270	0.294	0.920
		-0.5	0.005	0.213	0.222	0.943	0.022	0.292	0.308	0.930
	0.5	0.5	0.004	0.181	0.185	0.936	0	0.247	0.263	0.926
		0	-0.004	0.194	0.201	0.943	-0.008	0.256	0.278	0.915
		-0.5	0.007	0.215	0.217	0.943	0.008	0.275	0.291	0.933
	0.8	0.5	0.003	0.179	0.189	0.940	-0.003	0.209	0.233	0.916
		0	0.001	0.194	0.200	0.943	-0.003	0.222	0.244	0.923
		-0.5	0.003	0.214	0.226	0.936	0.010	0.241	0.267	0.921

Table 4. Estimation of β based on type II mixed data with p_r dependent on the covariate

			Proposed procedure				Zhu et al. (2013)			
$N_i^*(t)$	p_r	β_0	BIAS	ESE	SSE	CP	BIAS	ESE	SSE	CP
Poisson	0.2	0.5	0.006	0.144	0.154	0.939	0.010	0.224	0.236	0.934
		0	-0.001	0.161	0.170	0.932	-0.012	0.239	0.252	0.927
		-0.5	0	0.187	0.189	0.944	0.005	0.260	0.275	0.929
	0.5	0.5	0.005	0.144	0.150	0.945	0.007	0.212	0.229	0.927
		0	-0.001	0.161	0.162	0.943	0.006	0.223	0.241	0.927
		-0.5	-0.009	0.187	0.196	0.941	-0.007	0.243	0.260	0.933
	0.8	0.5	0.002	0.145	0.143	0.941	0.004	0.178	0.187	0.932
		0	0	0.162	0.169	0.936	0	0.189	0.210	0.917
		-0.5	-0.014	0.186	0.181	0.951	-0.009	0.211	0.213	0.942
Mixed	0.2	0.5	-0.001	0.179	0.187	0.937	-0.001	0.253	0.260	0.946
Poisson		0	0.004	0.192	0.196	0.941	0.002	0.263	0.281	0.933
		-0.5	-0.011	0.213	0.221	0.949	0	0.283	0.310	0.923
	0.5	0.5	-0.004	0.178	0.188	0.935	-0.002	0.238	0.254	0.931
		0	0.008	0.194	0.194	0.948	0.002	0.250	0.264	0.931
		-0.5	-0.008	0.215	0.220	0.949	-0.015	0.268	0.289	0.937
	0.8	0.5	-0.008	0.179	0.188	0.942	-0.006	0.208	0.230	0.915
		0	0.005	0.194	0.205	0.937	0.008	0.218	0.240	0.922
		-0.5	-0.016	0.215	0.221	0.949	-0.012	0.237	0.255	0.923

numerical study suggested that it still works without the assumption. As mentioned in previous sections, a main advantage of the proposed method is its easy implementation, as one can use the existing inference procedures and software packages for complete data. The simulation study indicated that the proposed methodology performs well for practical situations and is more efficient than the existing approach given in [3].

There exist several directions for future research. One is that for simplicity, we only considered the situation where both the observation process $O_i^*(t)$ and the follow-up time C_i^* are independent of covariates, and it is apparent that these may not be true in practice. Hence it will be useful to generalize the proposed estimation procedure to these situations. Another assumption behind the proposed method is the proportional rates model (1) and it is well-known that it may not fit the observed data well sometimes [1]. To address this, one may consider some other models such as the additive rates model or semiparametric transformation model [18, 19] and develop appropriate and valid estimation procedures. A more complicated situation could be that the observation process is informative [18], meaning that $O_i^*(t)$

is related to $N_i^*(t)$, even given covariates. In this case, the method given above is clearly not valid and thus one needs some other methods. In addition, it is clear that it would be helpful to derive or provide the theoretical justification for the normal approximation to the distribution of or the asymptotic normality of the proposed estimator $\hat{\boldsymbol{\beta}}$ as well as the asymptotic properties of $\hat{\Lambda}_0(t)$.

ACKNOWLEDGMENTS

The authors wish to thank the Editor, Dr. Heping Zhang, the Associate Editor and two referees for their helpful comments and suggestions that greatly improved the paper. The work was partly supported by NIH Grant (R03 CA169150; R21 CA198641) to Zhu and the funding from ALSAC and Cancer Center Support.

Received 16 September 2016

REFERENCES

[1] Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media. MR2597020

[2] Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. Springer, New York. MR3136574

[3] Zhu, L., Tong, X., Zhao, H., Sun, J., Srivastava, D. K., Leisenring, W. and Robison, L. L. (2013). Statistical analysis of mixed recurrent event data with application to cancer survivor study. *Statistics in Medicine* **32** 1954–1963. MR3067372

[4] Robison, L. L., Mertens, A. C., Boice, J. D., Breslow, N. E., Donaldson, S. S., Green, D. M., Li, F. P., Meadows, A. T., Mulvihill, J. J., Neglia, J. P., Nesbit, M. E., Packer, R. J., Potter, J. D., Sklar, C. A., Smith, M. A., Stovall, M., Strong, L. C., Yasui, Y. and Zeltzer, L. K. (2002). Study design and cohort characteristics of the childhood cancer survivor study: A multi-institutional collaborative project. *Medical and Pediatric Oncology* **38** 229–239.

[5] Zhu, L., Tong, X., Sun, J., Chen, M., Srivastava, D. K., Leisenring, W. and Robison, L. L. (2014). Regression analysis of mixed recurrent-event and panel-count data. *Biostatistics* **15** 555–568.

[6] Zhu, L., Zhao, H., Sun, J., Leisenring, W. and Robison, L. L. (2015). Regression analysis of mixed recurrent-event and panel-count data with additive rate models. *Biometrics* **71** 71–79. MR3335351

[7] Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*, volume 81. Wiley, New York. MR2117498

[8] Chen, L. and Sun, J. (2009). A multiple imputation approach to the analysis of current status data with the additive hazards model. *Communications in Statistics—Theory and Methods* **38** 1009–1018. MR2522544

[9] Chen, L. and Sun, J. (2010). A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. *Computational Statistics & Data Analysis* **54** 1109–1116. MR2580942

[10] Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd ed. Weily, Hoboken, New Jersey. MR1925014

[11] Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56** 199–203.

[12] Satten, G. A., Datta, S. and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93**, 318–327. MR1614581

[13] Tanner, M. A. and Wong, W. H. (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics* **29** 23–32.

[14] Wei, G. C. and Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics* 1297–1309.

[15] Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37** 158–168. MR1333194

[16] Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** 711–730. MR1796287

[17] Ross, S. M. (1983). *Stochastic Processes*. Wiley, New York. MR0683455

[18] Li, N., Zhao, H. and Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. *Statistics in Medicine* **32** 3039–3054. MR3073834

[19] Lin, D. Y., Wei, L. J. and Ying, Z. (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association* **96** 620–628. MR1946429

Guanglei Yu
Department of Statistics
University of Missouri-Columbia
Columbia, MO
65211
USA
E-mail address: gyb5b@mail.missouri.edu

Liang Zhu
Biostatistics and Epidemiology Research Design
University of Texas Health Science Center at Houston
Houston, TX
77030
USA
E-mail address: liang.zhu@uth.tmc.edu

Jianguo Sun
Department of Statistics
University of Missouri-Columbia
Columbia, MO
65211
USA
E-mail address: sunj@missouri.edu

Leslie L. Robison
Department of Epidemiology and Cancer Control
St. Jude Children's Research Hospital
Memphis, TN
38105
USA
E-mail address: les.robison@stjude.org