Published in final edited form as:

Scand Stat Theory Appl. 2019 June ; 46(2): 414–431. doi:10.1111/sjos.12357.

An Additive-Multiplicative Mean Model for Panel Count Data with Dependent Observation and Dropout Processes

Guanglei Yu¹, Yang Li², Liang Zhu³, Hui Zhao⁴, Jianguo Sun⁵, and Leslie L. Robison⁶

¹department of Statistics, University of Missouri-Columbia

²Department of Mathematics and Statistics, The University of North Carolina at Charlotte

³Division of Clinical and Translational Sciences, Department of Internal Medicine, University of Texas Health Science Center at Houston

⁴School of Statistics and Mathematics, Zhongnan University of Economics and Law, China

⁵Department of Statistics, University of Missouri-Columbia

⁶Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital

Abstract

This paper discusses regression analysis of panel count data with dependent observation and dropout processes. For the problem, a general mean model is presented that can allow both additive and multiplicative effects of covariates on the underlying point process. In addition, the proportional rates model and the accelerated failure time model are employed to describe possible covariate effects on the observation process and the dropout or follow-up process, respectively. For estimation of regression parameters, some estimating equation-based procedures are developed and the asymptotic properties of the proposed estimators are established. In addition, a resampling approach is proposed for estimating covariance matrix of the proposed estimator and a model checking procedure is also provided. Results from an extensive simulation study indicate that the proposed methodology works well for practical situations and it is applied to a motivating set of real data.

Keywords

Additive-multiplicative mean model; artificial censoring; dependent observation process; estimating equation; panel count data

1 | INTRODUCTION

This paper discusses regression analysis of panel count data that often arise from event history studies concerning the occurrence rates of certain recurrent events such as repeated disease infections, hospitalizations and warranty claims of automobiles. Note that for such studies, there usually exist two observation schemes. One is that every study subject is

observed continuously and the other is that each subject is observed only at discrete time points. For the former, complete data, which are commonly referred to as recurrent event data, are available for the underlying recurrent event process (Cook & Lawless, 2007), while the latter only yields incomplete data or so-called panel count data to be discussed in this paper (Sun & Zhao, 2013). In particular, instead of the exact event times, panel count data only provide the numbers of event occurrences between observation times.

An example of panel count data that motivated this study arose from the Childhood Cancer Survivor Study (CCSS) conducted by St. Jude Children's Research Hospital for evaluating long-term health effects of childhood cancer and cancer treatments (Zhu *et al.*, 2013; Robison *et al.*, 2002). This study consists of a cohort of childhood cancer survivors who were treated at St. Jude Children's Research Hospital and had survived at least 5 years, and various information was collected through questionnaires sent to the study subjects periodically. In the study, among others, one specific objective is to assess the effect or influence of health insurance plans and income levels on hospitalization rates. More details about the study will be given in Section 7. Many methods have been developed to analyze the panel count data (Sun & Zhao, 2013) and some early work includes Sun & Kalbfleisch (1995) and Wellner & Zhang (2000), who investigated nonparametric estimation for mean functions of an underlying recurrent event process based on panel count data. Also Sun & Wei (2000) and Lu *et al.* (2009) investigated regression analysis of panel count data and gave some estimating equation-based and likelihood-based approaches, respectively. However, one drawback of these methods is that they can only deal with the situation where observation processes are independent of an underlying recurrent event process given covariates.

Many authors have pointed out that sometimes a recurrent event process and its observation process may be related even given covariates (Li *et al.*, 2010; Sun & Zhao, 2013; Zhao & Tong, 2011). Among others, for example, Zhao & Tong (2011) and Deng (2013) discussed this issue and proposed some joint modeling approaches with the use of some latent variables to characterize the correlation. Also for the same problem, Buzkova (2010) and Li *et al.* (2010) gave some marginal modeling approaches, whose main feature is to leave the correlation structure unspecified. In addition to the possible correlation between the two processes above, sometimes we may also face the situation where the dropout or follow-up processes are also related to both the underlying recurrent event process and observation processes. Among others, both He *et al.* (2009) and Zhao *et al.* (2013) investigated the analysis of such panel count data and proposed some joint and marginal modeling procedures, respectively.

Note that all of the existing methods mentioned above assume that the effects of covariates are either additive or multiplicative, but as pointed out by Dong & Sun (2015), there may exist situations when both additive and multiplicative effects coexist. In the following, we will consider such situations and present an additive-multiplicative mean model for regression analysis of panel count data in the presence of both dependent observation and dropout processes. One feature or significance of the presented models is that they require no correlation structures or distribution assumptions on some latent processes in order to connect the three underlying related processes. This is a main advantage because it is well-

known that in practice, it is almost impossible to validate such distribution assumptions when data are incomplete as panel count data and also the inappropriate application of a distribution-assumption based procedure can lead to biased conclusions. Another feature of the proposed method is the use of the artificial censoring technique for the construction of estimating equations although it along with other features also introduces some techniques complications as seen in the Appendix.

The remainder of the paper will be organized as follows. We begin in Section 2 to describe the assumptions and models to be used throughout the paper and in particular, an additive-multiplicative mean model will be presented for the underlying recurrent event process. Furthermore, the proportional rates model and the accelerated failure time model will be employed to characterize covariate effects on the observation and dropout processes, respectively. In Section 3, we will first consider the situation when the underlying recurrent event process and its observation process may be correlated but both are independent of the dropout process given covariates. The proposed estimation procedure will then be generalized to the situation where all three processes can be correlated in Section 4. Section 5 discusses and presents a procedure for model checking and Section 6 reports some results from a simulation study conducted for assessing the finite-sample performance of the proposed approaches. In Section 7, we apply the methodology to the CCSS discussed above and Section 8 contains some discussion and concluding remarks.

2 | ASSUMPTIONS AND MODELS

Consider a recurrent event study consisting of n independent subjects. For subject i ($i = 1, \dots, n$), let $Y_i(t)$ denote the total number of event occurrences up to time t . Suppose that $Y_i(t)$ is observed at discrete time points $\{T_{i,1}, \dots, T_{i,m_i}\}$ only, where m_i represents the total number of

observations on subject i . Define the observation process by $N_i(t) = \sum_{j=1}^{m_i} I(T_{i,j} \leq t)$, which indicates the total number of observation times up to t . Also let $X_i(t)$ and $Z_i(t)$ denote two vectors of distinct external covariates that may depend on time and have possible additive and multiplicative effects on the mean function of $Y_i(t)$, respectively. More comments on this are given below. Also suppose that there exist two follow-up times D_i and C_i . Here we will assume that D_i denotes a dropout time that may be related to $Y_i(t)$ and $N_i(t)$, but C_i is independent of $\{Y_i(t), N_i(t), D_i\}$ given $\{Z_i(t), X_i(t)\}$. In reality, one observes $T_i = C_i \wedge D_i$, where $a \wedge b = \min(a, b)$. Denote $\delta_i = I(D_i < C_i)$.

To describe covariate effects, assume that there exist two correlated latent or error processes $\epsilon_{i1}(t)$ and $\epsilon_{i2}(t)$ such that given $X_i(t)$, $Z_i(t)$ and $\epsilon_{i1}(t)$, the mean function of $Y_i(t)$ satisfies the model

$$E\{Y_i(t) | X_i(t), Z_i(t), \epsilon_{i1}(t)\} = \{\beta'_0 X_i(t) + A_0(t; \epsilon_{i1}(t))\} e^{\theta'_0 Z_i(t)}. \quad (2.1)$$

In the above, β_0 and θ_0 are $p \times 1$ and $q \times 1$ vectors of regression parameters, respectively, and $A_0(t; \epsilon_{i1}(t))$ is a subject-specific baseline mean function nondecreasing on t with

$A_0(0; \epsilon_{i1}(0)) = 0$. As mentioned above, some special cases of or similar models to the model above have been discussed in literature and in particular, Liu *et al.* (2013) and Dong & Sun (2015) investigated similar models for regression analysis of recurrent event data.

For observation processes, we will assume that given $X_i(t)$, $Z_i(t)$, and $\epsilon_{i2}(t)$, $N_i(t)$ follows the proportional rates model given by

$$E\{dN_i(t) | X_i(t), Z_i(t), \epsilon_{i2}(t)\} = e^{\gamma_0' V_i(t)} dr_0(t; \epsilon_{i2}(t)). \quad (2.2)$$

Here $V_i(t)$ is a bounded subset or function of $Z_i(t)$ and $X_i(t)$, γ_0 a vector of unknown regression parameters, and $r_0(t; \epsilon_{i2}(t))$ an unknown subject-specific baseline mean function nondecreasing on t with $r_0(0; \epsilon_{i2}(0)) = 0$. It is apparent that correlation between $Y_i(t)$ and $N_i(t)$ is characterized by that between $\epsilon_{i1}(t)$ and $\epsilon_{i2}(t)$ and independence between the latter two processes means independence between the former two.

As mentioned above, the dropout time D_i can be related to $Y_i(t)$, $N_i(t)$ and covariates. To describe this, it will be assumed that D_i follows the accelerated failure time model defined as

$$D_i e^{-\eta_0' W_i} = D_0 \quad (2.3)$$

given covariates (Lin *et al.*, 1998; Tsiatis, 1990). In the above, as V_i , W_i is a bounded subset or function of $Z_i(t)$ and $X_i(t)$. η_0 is a vector of regression parameters, and D_0 is a failure time variable with unknown hazard function $\lambda_0(t)$. In the following, we will assume that joint distributions of $\{\epsilon_{i1}(t), \epsilon_{i2}(t), D_0\}_{i=1}^n$ are identical and unspecified. In other words, we leave the correlation structure among $\epsilon_{i1}(t)$, $\epsilon_{i2}(t)$ and D_0 to be arbitrary. Also we will assume that given $\{\epsilon_{i1}(t), \epsilon_{i2}(t), D_0, X_i(t), Z_i(t)\}$, $Y_i(t)$ and $N_i(t)$ are independent. Note that here for simplicity, we have assumed that the covariates W_i are time-independent. However, the estimation procedures proposed below can be easily generalized to the situation when D_i depends on some time-dependent covariates (Lin & Ying, 2003).

In the next section, we will first discuss the situation when the D_i 's are independent of the $Y_i(t)$'s and the $N_i(t)$'s given covariates and present an estimation procedure for all parameters of interest. The method will then be generalized to the situation when all three processes can be correlated.

3 | ESTIMATION WITH INDEPENDENT DROPOUT PROCESSES

To better describe the idea, in this section, we start with a simple scenario that only independent or noninformative dropouts are observed given covariates. That is, $C_i = D_i$ for all i . For estimation, define $\mu_0(t)$ and $\Lambda_0(t)$ by

$$E\left[\left[Y_i(t)e^{-\theta_0' Z_i(t)} - \beta_0' X_i(t)\right]e^{-\gamma_0' V_i(t)} dN_i(t)\right] = d\mu_0(t)$$

and

$$E \left\{ e^{-\gamma'_0 V_i(t)} dN_i(t) \right\} = d\Lambda_0(t).$$

Also define

$$M_i(t) = \int_0^t I(s \leq C_i) \left[\left\{ Y_i(s) e^{-\theta'_0 Z_i(s)} - \beta'_0 X_i(s) \right\} e^{-\gamma'_0 V_i(s)} dN_i(s) - d\mu_0(s) \right]$$

and

$$\mathcal{M}_i(t) = \int_0^t I(s \leq C_i) \left[dN_i(s) - e^{\gamma'_0 V_i(s)} d\Lambda_0(s) \right].$$

Then it is easy to show that $M_i(t)$ and $\mathcal{M}_i(t)$ are zero-mean stochastic processes and thus to estimate β_0 , θ_0 and $\mu_0(t)$, it is natural to consider the following estimating equations

$$\sum_{i=1}^n I(t \leq C_i) \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) - d\mu_0(t) \right] = 0, \quad (3.1)$$

$$\sum_{i=1}^n \int_0^\tau H(t) X_i(t) I(t \leq C_i) \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) - d\mu_0(t) \right] = 0, \quad (3.2)$$

and

$$\sum_{i=1}^n \int_0^\tau H(t) Z_i(t) I(t \leq C_i) \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) - d\mu_0(t) \right] = 0 \quad (3.3)$$

for given γ_0 . In the above, $H(t)$ is a possibly data-dependent weight function and τ denotes the length of study. Let $\zeta_0 = (\theta'_0, \beta'_0)'$ and $\zeta = (\theta', \beta')'$. By solving $\mu_0(t)$ from Equation (3.1) plugged into Equations (3.2) and (3.3), we obtain the estimating equation $U_1(\zeta; \gamma_0) = 0$ for ζ as

$$U_1(\zeta; \gamma_0) = \sum_{i=1}^n \int_0^\tau H(t) I(t \leq C_i) \left[X_i^*(t) - \bar{X}^*(t) \right] \left\{ Y_i(t) e^{-\theta' Z_i(t)} - \beta' X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t),$$

where $X_i^*(t) = (X'_i(t), Z'_i(t))'$ and $\bar{X}^*(t) = (\bar{X}'(t), \bar{Z}'(t))'$ with

$$\bar{X}(t) = \frac{\sum_{i=1}^n X_i(t) I(t \leq C_i)}{\sum_{i=1}^n I(t \leq C_i)} \text{ and } \bar{Z}(t) = \frac{\sum_{i=1}^n Z_i(t) I(t \leq C_i)}{\sum_{i=1}^n I(t \leq C_i)}.$$

Note that the equations above involve γ_0 , which is also usually unknown. On the other hand, it can be readily estimated from the following estimating equation

$$U_2(\gamma) = \sum_{i=1}^n \int_0^\tau \left\{ V_i(t) - \bar{V}_\gamma(t; \gamma) \right\} I(t \leq C_i) dN_i(t) = 0,$$

where

$$\bar{V}_\gamma(t; \gamma) = \frac{\sum_{i=1}^n I(t \leq C_i) e^{\gamma' V_i(t)} V_i(t)}{\sum_{i=1}^n I(t \leq C_i) e^{\gamma' V_i(t)}}.$$

Let $\hat{\gamma}_1$ denote the resulting estimator of γ_0 ; $\hat{\theta}_1$, $\hat{\beta}_1$ and $\hat{\mu}_0(t)$ denote the estimators of θ_0 , β_0 and $\mu_0(t)$, respectively, solved from Equations (3.1)–(3.3) with γ_0 replaced by $\hat{\gamma}_1$. Denote $\hat{\xi}_1 = (\hat{\theta}_1', \hat{\beta}_1')'$, and define

$$\hat{M}_i(t) = \int_0^t I(s \leq C_i) \left[\left(Y_i(s) e^{-\hat{\theta}_1' Z_i(s)} - \hat{\beta}_1' X_i(s) \right) e^{-\hat{\gamma}_1' V_i(s)} dN_i(s) - d\hat{\mu}_0(s) \right],$$

and

$$\hat{\mathcal{M}}_i(t) = \int_0^t I(s \leq C_i) \left[dN_i(s) - e^{\hat{\gamma}_1' V_i(s)} d\hat{\Lambda}_0(s) \right],$$

where

$$d\hat{\mu}_0(t) = \frac{\sum_{i=1}^n I(t \leq C_i) \left(Y_i(t) e^{-\hat{\theta}_1' Z_i(t)} - \hat{\beta}_1' X_i(t) \right) e^{-\hat{\gamma}_1' V_i(t)} dN_i(t)}{\sum_{i=1}^n I(t \leq C_i)},$$

and

$$d\hat{\Lambda}_0(t) = \frac{\sum_{i=1}^n I(t \leq C_i) dN_i(t)}{\sum_{i=1}^n I(t \leq C_i) e^{\hat{\gamma}_1' V_i(t)}}.$$

The theorem below establishes the asymptotic properties of $\hat{\xi}_1$.

Theorem 1. Assume that the regularity conditions (C1) - (C6) given in the Appendix hold. Then as $n \rightarrow \infty$, $\hat{\zeta}_1$ is a consistent estimator of ζ_0 and $n^{1/2}(\hat{\zeta}_1 - \zeta_0)$ converges in distribution to the normal random variable with mean zero and the covariance matrix that can be consistently estimated by $\hat{\Sigma}_\zeta = A_C^{-1} \hat{\Sigma} A_C'^{-1}$, where

$$A_C = \begin{pmatrix} A_C^{11} & A_C^{12} \\ A_C^{21} & A_C^{22} \end{pmatrix}, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\hat{u}_{1i} - \hat{u}_{2i}) \otimes^2.$$

In the above,

$$A_C^{11} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau H(t) (X_i(t) - \bar{X}(t)) I(t \leq C_i) Y_i(t) e^{-\hat{\theta}_1' Z_i(t)} Z_i'(t) e^{-\hat{\gamma}_1' V_i(t)} dN_i(t),$$

$$A_C^{12} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau H(t) (X_i(t) - \bar{X}(t)) I(t \leq C_i) X_i'(t) e^{-\hat{\theta}_1' Z_i(t)} dN_i(t),$$

$$A_C^{21} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau H(t) (Z_i(t) - \bar{Z}(t)) I(t \leq C_i) Y_i(t) e^{-\hat{\theta}_1' Z_i(t)} Z_i'(t) e^{-\hat{\gamma}_1' V_i(t)} dN_i(t),$$

$$A_C^{22} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau H(t) (Z_i(t) - \bar{Z}(t)) I(t \leq C_i) X_i'(t) e^{-\hat{\theta}_1' Z_i(t)} dN_i(t),$$

$$\hat{u}_{1i} = \int_0^\tau H(t) (X_i^*(t) - \bar{X}^*(t)) d\widehat{M}_i(t),$$

and

$$\hat{u}_{2i} = B \Omega^{-1} \int_0^\tau (V_i(t) - \bar{V}_\gamma(t; \hat{\gamma}_1)) d\widehat{\mathcal{M}}_i(t),$$

where

$$B = \frac{1}{n} \sum_{i=1}^n \int_0^\tau H(t) I(t \leq C_i) \left\{ X_i^*(t) - \bar{X}^*(t) \right\} \left\{ Y_i(t) e^{-\hat{\theta}_1' Z_i(t)} - \hat{\beta}_1' X_i(t) \right\} e^{-\hat{\gamma}_1' V_i(t)} V_i'(t) dN_i(t),$$

and

$$\Omega = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ V_i(t) - \bar{V}_\gamma(t; \hat{\gamma}_1) \right\} \otimes e^{\hat{\gamma}_1' V_i(t)} I(t \leq C_i) d\hat{\Lambda}_0(t)$$

With $a \otimes^2 = aa'$.

The proof of the result above is sketched in the Appendix. In the next section, we will generalize the estimation procedure above to a more general situation.

4 | ESTIMATION WITH DEPENDENT DROPOUT PROCESSES

Now we consider a more general situation where the dropout process may be related to both the underlying recurrent event and observation processes. To estimate β_0 and θ_0 , following a similar idea discussed above, define $\mu_0^*(t)$ and $\Lambda_0^*(t)$ by

$$E \left[\left\{ Y_i(t) e^{-\theta_0' Z_i(t)} - \beta_0' X_i(t) \right\} e^{-\gamma_0' V_i(t)} dN_i(t) | D_i e^{-\eta_0' W_i + d_0} \geq t \right] = d\mu_0^*(t)$$

and

$$E \left[e^{-\gamma_0' V_i(t)} dN_i(t) | D_i e^{-\eta_0' W_i + d_0} \geq t \right] = d\Lambda_0^*(t),$$

respectively, where $d_0 = \min_{1 \leq i \leq n} \eta_0' W_i$. Note that here we adopt the technique of artificial censoring (Lin & Ying, 2003): Dropout times D_i 's are replaced by potentially smaller, artificially created times $D_i e^{-\eta_0' W_i + d_0}$, which are stochastically equal for all i under the accelerated failure time model (2.3).

Furthermore, define $\xi_i(t; \eta_0) = I \left(T_i e^{-\eta_0' W_i + d_0} \geq t \right)$,

$$M_i^*(t) = \int_0^t \xi_i(s; \eta_0) \left[\left\{ Y_i(s) e^{-\theta_0' Z_i(s)} - \beta_0' X_i(s) \right\} e^{-\gamma_0' V_i(s)} dN_i(s) - d\mu_0^*(s) \right],$$

and

$$\mathcal{M}_i^*(t) = \int_0^t \xi_i(s; \eta_0) \left[dN_i(s) - e^{\gamma_0' V_i(s)} d\Lambda_0^*(s) \right].$$

Then we have

$$\begin{aligned}
E\{M_i^*(t)\} &= E\left\{\int_0^t \xi_i(s; \eta_0) \left[\left\{ Y_i(s) e^{-\theta'_0 Z_i(s)} - \beta'_0 X_i(s) \right\} e^{-\gamma'_0 V_i(s)} dN_i(s) - d\mu_0^*(s) \right] \right\} \\
&= E\left\{\int_0^t \xi_i(s; \eta_0) \left[E\left\{ \left\{ Y_i(s) e^{-\theta'_0 Z_i(s)} - \beta'_0 X_i(s) \right\} e^{-\gamma'_0 V_i(s)} dN_i(s) \mid T_i \leq s \right\} - d\mu_0^*(s) \right] \right\} \\
&= E\left\{\int_0^t \xi_i(s; \eta_0) \left[E\left\{ \left\{ Y_i(s) e^{-\theta'_0 Z_i(s)} - \beta'_0 X_i(s) \right\} e^{-\gamma'_0 V_i(s)} dN_i(s) \mid D_i \leq s \right\} - d\mu_0^*(s) \right] \right\} \\
&= E\left\{\int_0^t \xi_i(s; \eta_0) [d\mu_0^*(s) - d\mu_0^*(s)]\right\} = 0.
\end{aligned}$$

That is, $M_i^*(t)$ is a zero-mean stochastic process.

In the same manner, one can show that $E\{\mathcal{M}_i^*(t)\} = 0$ too. These naturally suggest that if η_0 and γ_0 are known, one can estimate β_0 , θ_0 and $\mu_0^*(t)$ by solving the following estimating equations

$$\sum_{i=1}^n \xi_i(t; \eta_0) \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) - d\mu_0^*(t) \right] = 0, \quad (4.1)$$

$$\sum_{i=1}^n \int_0^\tau H(t) X_i(t) \xi_i(t; \eta_0) \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) - d\mu_0^*(t) \right] = 0, \quad (4.2)$$

and

$$\sum_{i=1}^n \int_0^\tau H(t) Z_i(t) \xi_i(t; \eta_0) \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) - d\mu_0^*(t) \right] = 0. \quad (4.3)$$

Here as above, $H(t)$ denotes a possibly data-dependent weight function. By solving Equation (4.1) for $\mu_0^*(t)$ plugged into Equations (4.2) and (4.3), we obtain $U_3(\zeta; \gamma_0, \eta_0) = 0$ for ζ , where

$$U_3(\zeta; \gamma_0, \eta_0) = \sum_{i=1}^n \int_0^\tau H(t) \xi_i(t; \eta_0) \{X_i^*(t) - \bar{X}^*(t; \eta_0)\} \left[\left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) \right]$$

with $\bar{X}^*(t; \eta_0) = (\bar{X}'(t; \eta_0), \bar{Z}'(t; \eta_0))'$,

$$\bar{X}(t; \eta_0) = \frac{\sum_{i=1}^n X_i(t) \xi_i(t; \eta_0)}{\sum_{i=1}^n \xi_i(t; \eta_0)} \text{ and } \bar{Z}(t; \eta_0) = \frac{\sum_{i=1}^n Z_i(t) \xi_i(t; \eta_0)}{\sum_{i=1}^n \xi_i(t; \eta_0)}.$$

As in the previous section, γ_0 and η_0 are usually unknown and need to be estimated. For this purpose, define $N_i^D(t; \eta) = I\left(D_i e^{-\eta' W_i} \leq t\right)$, $\tilde{N}_i^D(t; \eta) = \delta_i I\left(T_i e^{-\eta' W_i} \leq t\right)$, and let $E\{dN_i^D(t; \eta)\} = d\Lambda_0^D(t)$. Also define

$$M_i^D(t) = \int_0^t d\tilde{N}_i^D(s; \eta_0) - I\left(T_i e^{-\eta_0' W_i} \geq s\right) d\Lambda_0^D(s).$$

Then it is easy to see that the $M_i^D(t)$'s are martingales and it follows that γ_0 and η_0 can be estimated from the following estimating equations

$$U_4(\gamma; \eta) = \sum_{i=1}^n \int_0^\tau \{V_i(t) - \bar{V}(t; \gamma, \eta)\} \xi_i(t; \eta) dN_i(t) = 0, \quad (4.4)$$

and

$$U_5(\eta) = \sum_{i=1}^n \int_0^\tau \{W_i - \bar{W}^D(t; \eta)\} d\tilde{N}_i^D(t; \eta) = 0, \quad (4.5)$$

where

$$\bar{V}(t; \gamma, \eta) = \frac{\sum_{i=1}^n \xi_i(t; \eta) e^{\gamma' V_i(t)} V_i(t)}{\sum_{i=1}^n \xi_i(t; \eta) e^{\gamma' V_i(t)}} \text{ and } \bar{W}^D(t; \eta) = \frac{\sum_{i=1}^n I\left(T_i e^{-\eta' W_i} \geq t\right) W_i}{\sum_{i=1}^n I\left(T_i e^{-\eta' W_i} \geq t\right)}.$$

Let $\hat{\gamma}_2$ and $\hat{\eta}$ denote the estimators of γ_0 and η_0 given by the solutions to the equations (4.4) and (4.5). Also let $\hat{\xi}_2 = (\hat{\theta}_2', \hat{\beta}_2')'$ and $\hat{\mu}_0^*(t)$ be the estimators of ζ_0 and $\hat{\mu}_0^*(t)$, respectively, given by the solutions to the equations (4.1)–(4.3) with γ_0 and η_0 replaced by $\hat{\gamma}_2$ and $\hat{\eta}$. The following theorem establishes the consistency and asymptotic normality of $\hat{\xi}_2$.

Theorem 2. Assume that the regularity conditions (C1) - (C7) given in the Appendix hold. Then as $n \rightarrow \infty$, $\hat{\xi}_2$ is a consistent estimator of ζ_0 and $n^{1/2}(\hat{\xi}_2 - \zeta_0)$ converges in distribution to the normal random variable with mean zero.

The proof of the result above is sketched in the Appendix. To apply it, it is apparent that we need to estimate the covariance matrix of $\hat{\xi}_2$, which is quite difficult or not practical. To address this, we propose to apply the resampling technique discussed in Parzen *et al.* (1994) as follows. Define

$$\widehat{M}_i^*(t; \hat{\xi}_2, \hat{\gamma}_2, \hat{\eta}) = \int_0^t \xi_i(s; \hat{\eta}) \left[Y_i(s) e^{-\hat{\theta}'_2 Z_i(s)} - \hat{\beta}'_2 X_i(s) \right] e^{-\hat{\gamma}'_2 V_i(s)} dN_i(s) - d\hat{\mu}_0^*(s),$$

$$\widehat{\mathcal{M}}_i^*(t; \hat{\gamma}_2, \hat{\eta}) = \int_0^t \xi_i(s; \hat{\eta}) \left[dN_i(s) - e^{\hat{\gamma}'_2 V_i(s)} d\widehat{\Lambda}_0^*(s) \right],$$

$$\widehat{M}_i^D(t; \hat{\eta}) = \int_0^t d\widehat{N}_i^D(s; \hat{\eta}) - I \left(T_i e^{-\hat{\eta}' W_i} \geq s \right) d\widehat{\Lambda}_0^D(s),$$

$$d\hat{\mu}_0^*(t) = \frac{\sum_{i=1}^n \xi_i(t; \hat{\eta}) \left[Y_i(t) e^{-\hat{\theta}'_2 Z_i(t)} - \hat{\beta}'_2 X_i(t) \right] e^{-\hat{\gamma}'_2 V_i(t)} dN_i(t)}{\sum_{i=1}^n \xi_i(t; \hat{\eta})},$$

$$d\widehat{\Lambda}_0^*(t) = \frac{\sum_{i=1}^n \xi_i(t; \hat{\eta}) dN_i(t)}{\sum_{i=1}^n \xi_i(t; \hat{\eta}) e^{\hat{\gamma}'_2 V_i(t)}}, \quad d\widehat{\Lambda}_0^D(t) = \frac{\sum_{i=1}^n d\widehat{N}_i^D(t; \hat{\eta})}{\sum_{i=1}^n I \left(T_i e^{-\hat{\eta}' W_i} \geq t \right)}.$$

Let $\widehat{\psi}_{1i} = \int_0^t H(t) \{X_i^*(t) - \bar{X}^*(t; \hat{\eta})\} d\widehat{M}_i^*(t; \hat{\xi}_2, \hat{\gamma}_2, \hat{\eta})$, $\widehat{\psi}_{i2} = \int_0^t \{V_i(t) - \bar{V}(t; \hat{\gamma}_2, \hat{\eta})\} d\widehat{\mathcal{M}}_i^*(t; \hat{\gamma}_2, \hat{\eta})$, $\widehat{\psi}_{i3} = \int_0^t \{W_i - \bar{W}(t; \hat{\eta})\} d\widehat{M}_i^D(t; \hat{\eta})$, and $(G_1, G_2, \dots, G_n)$ be independent standard normal variables independent of the observed data $\{Y(\cdot), N(\cdot), X(\cdot), Z(\cdot), T, \delta\}$. Furthermore let $\alpha = (\zeta', \gamma', \eta')'$ and $\hat{\alpha} = (\hat{\xi}'_2, \hat{\gamma}'_2, \hat{\eta}')'$. Also let $\alpha^* = (\zeta^{*'}, \gamma^{*'}, \eta^{*'})'$ denote the solutions of the following estimating equations

$$U_3(\zeta; \gamma, \eta) = - \sum_{i=1}^n \widehat{\psi}_{1i} G_i, \quad U_4(\gamma; \eta) = - \sum_{i=1}^n \widehat{\psi}_{2i} G_i, \quad U_5(\eta) = - \sum_{i=1}^n \widehat{\psi}_{3i} G_i.$$

Then it can be shown that the conditional distribution of $n^{1/2}(\zeta^* - \hat{\xi}_2)$ given the observed data is asymptotically equivalent to the unconditional distribution of $n^{1/2}(\hat{\xi}_2 - \zeta_0)$. Thus it follows that to approximate the distribution of $\hat{\xi}_2$, one can obtain a large number of realizations of ζ^* by generating standard normal random samples $(G_1, \dots, G_n)$ repeatedly while fixing the data $\{Y(\cdot), N(\cdot), X(\cdot), Z(\cdot), T, \delta\}$ at their observed values. In particular, we can estimate the asymptotic covariance matrix of $\hat{\xi}_2$ by the empirical covariance matrix of ζ^* .

5 | A MODEL CHECKING PROCEDURE

In practice, a natural question of interest is to check the appropriateness of models (2.1) - (2.3). Motivated by the ideas discussed in Lin *et al.* (1996) and Ghosh & Lin (2003), we consider the following three residual processes

$$\mathcal{F}_1(t; \alpha) = \sum_{i=1}^n \int_0^t X_i^*(s) d\widehat{M}_i^*(s; \alpha),$$

$$\mathcal{F}_2(t; \gamma, \eta) = \sum_{i=1}^n \int_0^t V_i(s) d\widehat{\mathcal{M}}_i^*(s; \gamma, \eta),$$

and

$$\mathcal{F}_3(t; \eta) = \sum_{i=1}^n W_i \widehat{M}_i^D(t; \eta).$$

Let $\mathcal{F}(t; \alpha) = (\mathcal{F}_1(t; \alpha)', \mathcal{F}_2(t; \gamma, \eta)', \mathcal{F}_3(t; \eta)')'$. Under models (2.1) - (2.3), $\mathcal{F}(t; \hat{\alpha})$ is expected to fluctuate randomly around 0. Thus it is reasonable to perform a goodness-of-fit test on these models based on the three supremum statistics $\sup_t \|\mathcal{F}_1(t; \hat{\alpha})\|$, $\sup_t \|\mathcal{F}_2(t; \hat{\gamma}_2, \hat{\eta})\|$ and $\sup_t \|\mathcal{F}_3(t; \hat{\eta})\|$.

For implementation of these tests and determination of the corresponding p -values, first note that the distribution of $\mathcal{F}(t; \hat{\alpha})$ can be approximated by the conditional distribution of $\mathcal{F}^*(t) = (\mathcal{F}_1^*(t)', \mathcal{F}_2^*(t)', \mathcal{F}_3^*(t)')'$, where

$$\mathcal{F}_1^*(t) = \sum_{i=1}^n \int_0^t \{X_i^*(s) - \bar{X}^*(s; \hat{\eta})\} d\widehat{M}_i^*(s; \hat{\alpha}) G_i + \mathcal{F}_1(t; \alpha^*) - \mathcal{F}_1(t; \hat{\alpha}),$$

$$\mathcal{F}_2^*(t) = \sum_{i=1}^n \int_0^t \{V_i(s) - \bar{V}(s; \hat{\gamma}_2, \hat{\eta})\} d\widehat{\mathcal{M}}_i^*(s; \hat{\gamma}_2, \hat{\eta}) G_i + \mathcal{F}_2(t; \gamma^*, \eta^*) - \mathcal{F}_2(t; \hat{\gamma}_2, \hat{\eta}),$$

and

$$\mathcal{F}_3^*(t) = \sum_{i=1}^n \int_0^t \{W_i - \bar{W}(s; \hat{\eta})\} d\widehat{M}_i^D(s; \hat{\eta}) G_i + \mathcal{F}_3(t; \eta^*) - \mathcal{F}_3(t; \hat{\eta})$$

with $(G_1, G_2, \dots, G_n)$ being independent standard normal random variables independent of the observed data. This suggests that for the goodness-of-fit test based on $\sup_t \|\mathcal{F}_1(t; \hat{\alpha})\|$,

$\sup_t \|\mathcal{F}_2(t; \hat{\gamma}_2, \hat{\eta})\|$ and $\sup_t \|\mathcal{F}_3(t; \hat{\eta})\|$, one can obtain p -values by comparing their observed values to the realizations generated by repeatedly generating a large number of standard normal random samples $(G_1, G_2, \dots, G_n)$ while fixing the observed data. Alternatively, one can perform a graphical model checking by plotting $\mathcal{F}(t; \hat{\alpha})$ to a few of realizations of $\mathcal{F}^*(t)$ and an unusual pattern of $\mathcal{F}(t; \hat{\alpha})$ suggests lack-of-fit of the models.

6 | A SIMULATION STUDY

In this section, we present some results obtained from an extensive simulation study conducted to assess the finite-sample performance of the procedures proposed in previous sections. In the study, we generated the covariate Z_i from the Bernoulli distribution with the success probability of 0.5. For covariate X_i , we first generated $\tilde{X}_i$ from the standard normal distribution and then set $X_i = \lfloor \tilde{X}_i \rfloor$. Also we set both $V_{\lambda}(t)$ and W_i as Z_i and assumed that non-informative follow-up times C_i 's followed the uniform distribution over $(\tau/2, \tau)$ with $\tau = 1.5$.

For the observed dropout times T_i 's, we considered both cases discussed in Sections 3 and 4. For the former, we set $T_i = C_i$ for $i = 1, \dots, n$ and then generated the random variables ϵ_i 's from a normal distribution with both mean and standard deviation being equal to 0.25. It is assumed that the underlying recurrent event process $Y_{\lambda}(t)$ and the observation process $N_{\lambda}(t)$ are mixed Poisson processes satisfying models (2.1) and (2.2) with $A_0(t, \epsilon_i(t)) = (2 + |\epsilon_i|) * t$ and $r_0(t, \epsilon_i(t)) = (3 + |\epsilon_i|) * t$, respectively. For the latter case, both observation times and panel count data were generated in the same way as above. In addition, we set $D_i = D_0 e^{\eta_0 W_i}$ with $D_0 = e^{\epsilon_i}$ and $T_i = C_i \wedge D_i$. In the following, the results are based on $n = 200, 1000$ replications and $H(t) = 1$.

Table 1 gives the results on the estimators $\hat{\beta}_1$ and $\hat{\theta}_1$ with the true values β_0, θ_0 and γ_0 all being 0 or 0.5. They include the estimated bias given by the average of each parameter estimates minus their true value (BIAS), the average of the estimated standard errors (ESE), the sample standard error of the estimates (SSE) and the 95% empirical coverage probabilities (CP). One can see from the table that the proposed estimator seems to be unbiased and the proposed variance estimation also appears reasonable. In addition, the coverage probabilities are all close to their nominal level, suggesting that the normal approximation to the distribution of the proposed estimator seems to be appropriate too.

Estimation results for $\hat{\beta}_2$ and $\hat{\theta}_2$ obtained with informative dropout times are presented in Table 2 with the same true values of β_0, θ_0 and γ_0 as for Table 1. In addition, we set $\eta_0 = 0$ or 0.2. Again the results indicate that the proposed estimation procedure performs well and the normal approximation to distribution of the proposed estimates appears reasonable. We also considered other set-ups and obtained similar results and conclusions. In particular, note that for the data generated above, the average number of recurrent events per subject ranges from 1 to 3. Corresponding to this, we simulated some data with the average numbers of the recurrent events ranging from 3 to 8 and obtained similar conclusions. Here we focused on

the estimation of β and θ only as the estimation of γ and η are relatively straightforward and have been discussed in literature.

By following the suggestion of a reviewer, we also conducted some simulation studies to evaluate the performance of the model checking procedure described in Section 5 with the focus on the statistic $\sup_t \|\mathcal{F}_1(t; \hat{\alpha})\|$ for Model (2.1). This is because the test procedures for models (2.2) and (2.3) have been discussed in the literature. In the study, we assumed that instead of model (2.1), the true model of $Y_i(t)$ took the form

$$E\{Y_i(t)|X_i(t), Z_i(t), \epsilon_{i1}(t)\} = \{\beta'_0 X_i(t) + A_0(t; \epsilon_{i1}(t))\} e^{\theta'_0 Z_i(t)} + k B_0(t) X_i(t)$$

with $k = 0, 0.5, 1, 1.5, 2, 2.5$ and $B_0(t) = t^2$. Table 3 reports the estimated test sizes ($k = 0$) and powers ($k \neq 0$) of the proposed model-checking procedure at the significant level 0.05 based on $\sup_t \|\mathcal{F}_1(t; \hat{\alpha})\|$ with $\eta_0 = 400$ or 800 and 1000 replications. Two scenarios are considered when $\beta_0 = 0$, $\theta_0 = 0$, $\gamma_0 = 0$, and $\eta_0 = 0$: one with the data generated in the same set-ups as those for Table 2 but $r_0(t; \epsilon_{i2}(t)) = 5 * t$ (Scenario 1), and the other with exactly the same set-ups as for Table 2 (Scenario 2). One can see that the results indicate that the proposed model checking procedure works well with the empirical sizes close to their nominal level and good powers to detect the model misspecification. As expected, the power increased as the deviation from the null model became more severe or when the sample size increased. Besides, also as expected, the test procedure had more powers for scenario 1 than for scenario 2.

7 | AN APPLICATION

Now we apply the methodology proposed in the previous sections to a set of panel count data arising from the CCSS discussed above, for which one primary aim is to characterize or investigate the late effects of childhood cancer on the health of the childhood cancer survivors. In the study, the participants were distributed a baseline summary questionnaire in 1996, and then followed up by summary questionnaires periodically thereafter including questions about chronic health conditions among others. Since the survivors are at risk for a myriad of late effects related to their cancer treatment, it is very important for them to have access to long-term follow-up care and cancer screening. However, the previous research (Park *et al.*, 2017) revealed that the childhood survivors are vulnerable to be uninsured and have low income, which could be key barriers to accessing preventive care to prevent or lessen future morbidity and mortality. To assess this, in the following, we will focus on the hospital visits and evaluate whether no insurance plan and low income (defined as household income $\leq$ \$19,999 per year) had some effects on the hospitalization rates. Such information can help to potentially impact affordable healthcare planning for these patients in terms of screening for and the management of chronic health conditions.

Among others, Kurt *et al.* (2012) analyzed the study and pointed out that Hodgkins disease survivors had the highest hospitalization rate among all childhood cancer diagnosis including leukemia, central nervous system tumors, non-Hodgkin's lymphoma, kidney

tumors, neuroblastoma, soft tissue sarcoma and bone cancer. In the following, we will confine us to the 1082 survivors of childhood Hodgkin's disease with the complete information on whether the subject was covered by health insurance and if he/she belonged to the low income group. Among them, the number of hospitalizations ranges from 0 to 34 with the average being 0.68, 897 subjects had health insurance and 258 belonged to the low income group. On the follow-up, 94 subjects died during the study with the median age being 43.5 years old, and each death time will be treated to be an informative dropout time D_i .

To apply the proposed methodology, let $Y_i(t)$ denote the total number of hospitalizations up to time t for subject i . Define $INS = 1$ if the subject was covered by health insurance and 0 otherwise, and $LOW = 1$ if the subject belonged to the low income group and 0 otherwise. For the model selection or determination of covariate types for the hospitalization process, we considered all of the following four possible models

$$AM1: E\left\{Y(t) | INS, LOW, \epsilon_1(t)\right\} = \{\beta_{01}LOW + A_0(t; \epsilon_1(t))\}e^{\theta_{01}INS},$$

$$AM2: E\left\{Y(t) | INS, LOW, \epsilon_1(t)\right\} = \{\beta_{02}INS + A_0(t; \epsilon_1(t))\}e^{\theta_{02}LOW},$$

$$MM: E\left\{Y(t) | INS, LOW, \epsilon_1(t)\right\} = A_0(t; \epsilon_1(t))e^{\theta_{01}INS + \theta_{02}LOW},$$

and

$$AA: E\left\{Y(t) | INS, LOW, \epsilon_1(t)\right\} = \beta_{01}LOW + \beta_{02}INS + A_0(t; \epsilon_1(t)).$$

Note that here we set $V_i(t)$ and W_i to be the same and include both covariates INS and LOW .

For the analysis, we will first see how each of the four models above fits the data along with models (2.2) and (2.3) for questionnaires return rates and dropout times by applying the model checking procedures described in Section 5. Table 4 gives the p -values obtained for testing each model and they suggest that all four candidate models with models (2.2) and (2.3) seem to fit the data well. To further investigate this, we applied the Akaike information criterion (AIC) defined as $AIC = 2k/n + \log(RSS/n)$, where $RSS = \sum_{i=1}^n \widehat{M}_i^*(\tau)^2$ and k denotes the number of parameters in the model. The AIC values for models AM1, AM2, MM, and AA are 0.721, 0.187, -0.700, and 1.676, respectively, and indicate that the candidate model MM should be used.

Table 5 presents the analysis results given by the estimation procedure described in Section 4 based on model MM, including the estimated covariate effects, the estimated standard

errors and the p -values for testing no covariate effect. They suggest that the hospitalization rate did not seem to be significantly related to either the income level of the patient or if the subject was covered by health insurance. Also whether a patient had health insurance did not seem to affect the return rate of the questionnaires or the mortality rate. On the other hand, the patients' income level seems to be significantly related to both rates, and the subjects with the lower income level tended to have both higher return rate of the questionnaires and higher risk of mortality than the other subjects. For comparison, we also applied the method given in Section 3 and that proposed by Hu *et al.* (2003) to the data and included the results in Table 5. Note that the former assumed that the hospitalization process may be related to the returning rate of the questionnaires but not the mortality, while the latter assumed that the hospitalization process is independent of both the return rate and the mortality. One can see from the table that both methods yielded similar results and they indicate that both covariates had significant effects on the return rate of the questionnaires. Also the results suggest that the income level was significantly related to the hospitalization rate. In other words, the analysis that ignores the existence of informative observation and follow-up processes could lead to different or misleading results and conclusions.

8 | DISCUSSION AND CONCLUDING REMARKS

This paper discussed regression analysis of panel count data in the presence of informative observation and dropout processes and for the problem, a general semiparametric regression model was presented. The proposed model includes many existing models as special cases and allows both multiplicative and additive covariate effects. In addition, some latent processes were employed to characterize the relationship among an underlying recurrent event process, the observation observation process and the dropout process. For estimation, estimating equation-based procedures were developed and the asymptotic properties of the resulting estimators were established. Also a resampling technique was applied for estimating covariance matrix of the estimators and a goodness-of-fit test procedure was provided for checking the appropriateness of the proposed models. The numerical study indicated that the proposed methodology performs well for practice situations and one could obtain misleading conclusions by ignoring the informative observation and dropout processes.

In the proposed methodology, we have used the proportional rates model and the accelerated failure time model for the observation and dropout processes, respectively. It is apparent that in practice, there may exist situations where one or both models may not be appropriate. To deal with them, for the former, an alternative is the additive rates or means model (Cook & Lawless, 2007), while for the latter, one may use the proportional hazards model or the additive hazards model instead. Of course, many other models are available for both processes such as some transformation types of models (Kalbfleisch & Prentice, 2011; Sun & Zhao, 2013). Under any combination of these models, one can apply the same idea discussed above to develop some estimating equation-based estimation procedures as well as goodness-of-fit test procedures although some details will be different.

Note that for determining the possible multiplicative or additive effect of a covariate, in Section 7, we suggested an AIC-based procedure by considering all possible combinations.

It is easy to see that this may not be feasible when there exists a large number of covariates. For this latter situation, one possible approach is to perform some initial data exploration or to check each covariate individually, not together, to see if it has multiplicative or additive effects by using, for example, a similar AIC-based or graphical method. Also in an application, one may employ the underlying biological or physical meaning or interpretations of a covariate to determine if it should be modeled multiplicatively or additively.

Also with the development of the proposed estimation procedures, we have applied the artificial censoring technique to correct the biases caused by the informative follow-ups. It should be noted that this could cause some efficiency loss when the sample size is very small and as an alternative, one may instead apply the inverse probability technique or other methods. The proposed estimator with the use of the artificial censoring technique could also be less efficient if the range of $\eta'_0 W$ is very large when, for example, there exist some extreme values for some components of W because it may cause excessive artificial censoring. To remedy this problem, as suggested in Lin & Ying (2003), one can alleviate this problem by stratifying the samples on W and constructing estimating functions separately for different strata. In consequence, the regression parameters can be estimated by solving the combined estimating equations together.

ACKNOWLEDGMENTS

The authors wish to thank the editors and the referees for their many helpful comments and suggestions that greatly improved the paper. The work was partly supported by NIH Grant (R21CA198641; R03CA219450) to Zhu and by the National Natural Science Foundation of China (Grant Nos. 11471135, 11571133) to Zhao.

Funding information

NIH Grant (R21CA198641; R03CA219450), National Natural Science Foundation of China (GrantNos. 11471135, 11571133)

Appendix: Asymptotic Normality of the Proposed Estimators

To prove the asymptotic normality of $\hat{\xi}_1$ and $\hat{\xi}_2$, we need the following regularity conditions:

(C1) $\{Y_i(\cdot), N_i(\cdot), X_i(\cdot), Z_i(\cdot), T_i, \delta_i\}_{i=1}^n$ are independent and identically distributed.

(C2) $P(T = \tau) > 0$, and $Y(\tau)$ is bounded by a constant.

(C3) The covariates $X(t)$ and $Z(t)$ have bounded variation on $[0, \tau]$.

(C4) The weight function $H(t)$ converges almost surely to a bounded and deterministic function $H(t)$ uniformly in $t \in [0, \tau]$.

(C5) Ω_0 is nonsingular, where $\Omega_0 = E \left\{ \int_0^\tau \left\{ V_i(t) - \bar{v}_{\gamma_0}(t) \right\}^{\otimes 2} e^{\gamma_0' V_i(t)} I(t \leq C_i) d\Lambda_0(t) \right\}$ and $\bar{v}_{\gamma_0}(t)$ is the limit of $\bar{v}_{\gamma_0}(t; \gamma_0)$.

(C6) A is nonsingular, where $A = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}$, where

$$A^{11} = E \left\{ \int_0^\tau h(t) (X_i(t) - \bar{x}(t)) I(t \leq C_i) Y_i(t) e^{-\theta'_0 Z_i(t)} Z_i(t) e^{-\gamma'_0 V_i(t)} dN_i(t) \right\},$$

$$A^{12} = E \left\{ \int_0^\tau h(t) (X_i(t) - \bar{x}(t)) I(t \leq C_i) X_i'(t) e^{-\gamma'_0 V_i(t)} dN_i(t) \right\},$$

$$A^{21} = E \left\{ \int_0^\tau h(t) (Z_i(t) - \bar{z}(t)) I(t \leq C_i) Y_i(t) e^{-\theta'_0 Z_i(t)} Z_i(t) e^{-\gamma'_0 V_i(t)} dN_i(t) \right\},$$

$$A^{22} = E \left\{ \int_0^\tau h(t) (Z_i(t) - \bar{z}(t)) I(t \leq C_i) X_i'(t) e^{-\gamma'_0 V_i(t)} dN_i(t) \right\},$$

and $\bar{x}(t)$, $\bar{z}(t)$ are limits of $\bar{X}(t)$, $\bar{Z}(t)$, respectively.

(C7) A_α , the asymptotic slope matrix of $n^{-1} U(\alpha_0)$, is nonsingular, where $U(\alpha_0) = (U'_3(\zeta_0; \gamma_0, \eta_0)', U'_4(\gamma_0; \eta_0)', U'_5(\eta_0)')'$.

The regularity conditions above are commonly used in the panel count data literature and can be satisfied in many situations. The conditions (C1) - (C2) are similar to those of Andersen & Gill (1982, Theorem 4.1) and are standard conditions for both panel count data and failure time data. The condition (C3) is about the boundedness and continuity, commonly used in survival analysis, while the condition (C4) requires that the weight function is convergent as n goes to infinity. The conditions (C5) - (C7) assume that the matrixes are nonsingular, and the conditions (C4) - (C7) are crucial for the existence and uniqueness of the proposed estimator to avoid technical distractions in the theorems and their proofs. Note that we do not need to assume $Y(t) > 0$.

First we prove the asymptotic normality of $\hat{\xi}_1$. For this, it follows from (3.1) that

$$d\hat{\mu}_0(t) = \frac{\sum_{i=1}^n I(t \leq C_i) \left[Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right] e^{-\gamma'_0 V_i(t)} dN_i(t)}{\sum_{i=1}^n I(t \leq C_i)}.$$

Plugging it into (3.2) and (3.3) with some simple algebra, we have

$$\sum_{i=1}^n \int_0^\tau h(t) I(t \leq C_i) [X_i(t) - \bar{X}(t)] \left[Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right] e^{-\gamma'_0 V_i(t)} dN_i(t) = 0$$

and

$$\sum_{i=1}^n \int_0^\tau H(t) I(t \leq C_i) [Z_i(t) - \bar{Z}(t)] \left\{ Y_i(t) e^{-\theta'_0 Z_i(t)} - \beta'_0 X_i(t) \right\} e^{-\gamma'_0 V_i(t)} dN_i(t) = 0.$$

Define

$$U_1(\zeta; \hat{\gamma}_1) = \sum_{i=1}^n \int_0^\tau H(t) I(t \leq C_i) [X_i^*(t) - \bar{X}^*(t)] \left\{ Y_i(t) e^{-\theta' Z_i(t)} - \beta' X_i(t) \right\} e^{-\hat{\gamma}'_1 V_i(t)} dN_i(t).$$

Then by the strong law of large numbers and multivariate central limit theorem, one can show that

$$\begin{aligned} n^{1/2}(\hat{\zeta}_1 - \zeta_0) &= \left[-n^{-1} \frac{\partial U_1(\zeta_0; \hat{\gamma}_1)}{\partial \zeta'} \right]_{\zeta = \zeta_0}^{-1} n^{-1/2} U_1(\zeta_0; \hat{\gamma}_1) + o_p(1) \\ &= A_c^{-1} \left[n^{-1/2} U_1(\zeta_0; \gamma_0) + n^{-1} \frac{\partial U_1(\zeta_0; \gamma_0)}{\partial \gamma'} \right]_{\gamma = \gamma_0} \left[-n^{-1} \frac{\partial U_2(\gamma_0)}{\partial \gamma'} \right]_{\gamma = \gamma_0}^{-1} n^{-1/2} U_2(\gamma_0) + o_p(1) \\ &= A_c^{-1} \left[n^{-1/2} \sum_{i=1}^n u_{1i} - B \Omega^{-1} n^{-1/2} \sum_{i=1}^n u_{2i} \right] + o_p(1), \end{aligned}$$

where $u_{1i} = \int_0^\tau h(t) \{X_i^*(t) - \bar{x}^*(t)\} dM_i(t)$ and $u_{2i} = \int_0^\tau \{V_i(t) - \bar{v}_{\gamma_0}(t)\} d\mathcal{M}_i(t)$. Thus the asymptotic normality property of $n^{1/2}(\hat{\zeta}_1 - \zeta_0)$ follows.

Now we prove the asymptotic normality of $\hat{\zeta}_1$. By following the similar arguments of Fleming & Harrington (2011) and Lin & Ying (2001), one can obtain

$$n^{-1/2} U_3(\zeta_0; \gamma_0, \eta_0) = n^{-1/2} \sum_{i=1}^n \int_0^\tau h(t) (X_i^*(t) - \bar{x}^*(t)) dM_i^*(t) + o_p(1),$$

$$n^{-1/2} U_4(\gamma_0; \eta_0) = n^{-1/2} \sum_{i=1}^n \int_0^\tau (V_i(t) - \bar{v}(t)) d\mathcal{M}_i^*(t) + o_p(1),$$

$$n^{-1/2} U_5(\eta_0) = n^{-1/2} \sum_{i=1}^n \int_0^\tau (W_i - \bar{w}^D(t)) dM_i^D(t) + o_p(1),$$

where $\bar{x}^*(t)$, $\bar{v}(t)$ and $\bar{w}^D(t)$ are limits of $\bar{X}^*(t; \eta_0)$, $\bar{V}(t; \gamma_0, \eta_0)$ and $\bar{W}^D(t; \eta_0)$ respectively. The multivariate central limit theorem implies that $n^{-1/2}U(\alpha_0)$ is asymptotically zero-mean normal with the covariance matrix

$$\Sigma_\alpha = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n E\{\psi_i \otimes^2\},$$

where $\psi_i = (\psi'_{1i}, \psi'_{2i}, \psi'_{3i})'$, $\psi_{1i} = \int_0^\tau h(t)(X_i^*(t) - \bar{x}^*(t))dM_i^*(t)$, $\psi_{2i} = \int_0^\tau (V_i(t) - \bar{v}(t))d\mathcal{M}_i^*(t)$, $\psi_{3i} = \int_0^\tau (W_i - \bar{w}^D(t))dM_i^D(t)$. By using the same arguments of Ying (1993), one can show that, for α in a small neighborhood of α_0 ,

$$n^{-1/2}U(\alpha) = n^{-1/2}U(\alpha_0) + A_\alpha n^{1/2}(\alpha - \alpha_0) + o_p(1).$$

Therefore, $n^{1/2}(\hat{\alpha} - \alpha_0)$ is asymptotically zero-mean normal with the covariance matrix $A_\alpha^{-1} \Sigma_\alpha A_\alpha'^{-1}$. This proves the asymptotic normality of $n^{1/2}(\hat{\zeta}_2 - \zeta_0)$.

Since $U(\alpha^*) = -\sum_{i=1}^n \psi_i G_i$, then

$$n^{1/2}(\alpha^* - \hat{\alpha}) = -A_\alpha^{-1} n^{-1/2} \sum_{i=1}^n \psi_i G_i + o_p(1).$$

It is straightforward to show the conditional distribution of $n^{-1/2} \sum_{i=1}^n \psi_i G_i$ given the observed data is zero-mean normal with covariance matrix $n^{-1} \sum_{i=1}^n \hat{\psi}_i \otimes^2$, which converges in probability to Σ_α . Thus, the conditional distribution of $n^{1/2}(\alpha^* - \hat{\alpha})$ is asymptotically equivalent to the unconditional distribution of $n^{1/2}(\hat{\alpha} - \alpha_0)$. In particular, one can use the conditional distribution of $n^{1/2}(\zeta^* - \hat{\zeta}_2)$ given the observed data to estimate the unconditional distribution of $n^{1/2}(\hat{\zeta}_2 - \zeta_0)$.

References

- Andersen PK, & Gill RD (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* 1100–1120.
- Buzkova P (2010). Panel count data regression with informative observation times. *Int. J. Biostat* 6, article 30.
- Cook RJ, & Lawless J (2007). *The statistical analysis of recurrent events*. Springer Science & Business Media.
- Deng S (2013). Semiparametric regression analysis of panel count data with time-dependent covariates and informative observation and censoring times. *Communications in Statistics Theory and Methods* 42, 4170–4183.
- Dong L, & Sun L (2015). A flexible semiparametric transformation model for recurrent event data. *Lifetime Data Anal.* 21, 20–41. [PubMed: 24241908]

- Fleming TR, & Harrington DP (2011). Counting processes and survival analysis. John Wiley & Sons.
- Ghosh D, & Lin DY (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics* 59, 877–885. [PubMed: 14969466]
- Han M, Song X, Sun L, & Liu L (2014). Joint modeling of longitudinal data with informative observation times and dropouts. *Statist. Sinica* 24, 1487–1504.
- He X, Tong X, & Sun J (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Anal.* 15, 177–196. [PubMed: 19082711]
- Hu XJ, Sun J, & Wei LJ (2003). Regression parameter estimation from panel counts. *Scand. J. Stat* 30, 25–43.
- Kalbfleisch JD, & Prentice RL (2011). The statistical analysis of failure time data. John Wiley & Sons: New York, 2011.
- Kurt BA, Nolan VG, Ness KK, Neglia JP, Tersak JM, Hudson MM, Armstrong GT, Hutchinson RJ, Leisenring MM, Oeffinger KC, Robison LL, & Arora M (2012). Hospitalization rates among survivors of childhood cancer in the Childhood Cancer Survivor Study cohort. *Pediatr Blood Cancer* 59, 126–132. [PubMed: 22180128]
- Li N, Sun L, & Sun J (2010). Semiparametric transformation models for panel count data with dependent observation process. *Statistics in Biosciences* 2, 191–210.
- Lin DY, & Ying Z (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc* 96, 103–126.
- Lin DY, & Ying Z (2003). Semiparametric regression analysis of longitudinal data with informative dropouts. *Biostatistics* 4, 385–398. [PubMed: 12925506]
- Lin DY, Robins JM, & Wei LJ (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* 83, 381–393.
- Lin DY, Wei LJ, & Ying Z (1998). Accelerated failure time models for counting processes. *Biometrika* 85, 605–618.
- Liu Y, Wu Y, Cai J, & Zhou H (2010). Additive-multiplicative rates model for recurrent events. *Lifetime Data Anal.* 16, 353–373. [PubMed: 20229314]
- Lu M, Zhang Y, & Huang J (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *J. Amer. Statist. Assoc* 104, 1060–1070.
- Park ER, Kirchhoff AC, Nipp RD, Donelan K, Leisenring WM, Armstrong GT, & Kuhlthau KA (2017). Assessing health insurance coverage characteristics and impact on health care cost, worry, and access: a report from the Childhood Cancer Survivor study. *JAMA internal medicine* 177, 1855–1858. [PubMed: 28973059]
- Parzen MI, Wei LJ, & Ying Z (1994). A resampling method based on pivotal estimating functions. *Biometrika* 81, 341–350.
- Robison LL, Mertens AC, Boice JD, Breslow NE, Donaldson SS, Green DM, Li FP, Meadows AT, Mulvihill JJ, Neglia JP, Nesbit ME, Packer RJ, Potter JD, Sklar CA, Smith MA, Stovall M, Strong LC, Yasui Y, & Zeltzer LK (2002). Study design and cohort characteristics of the childhood cancer survivor study: A multi-institutional collaborative project. *Medical and Pediatric Oncology* 38, 229–239. [PubMed: 11920786]
- Sun J, & Kalbfleisch JD (1995). Estimation of the mean function of point processes based on panel count data. *Statist. Sinica* 5, 279–289.
- Sun J, & Wei LJ (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B* 62, 293–302.
- Sun J, & Zhao X (2013). Statistical analysis of panel count data. New York: Springer.
- Tsiatis AA (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist* 18, 354–372.
- Wellner JA, & Zhang Y (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist* 28, 779–814.
- Ying Z (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist* 21, 76–99.
- Zhao H, Li Y, & Sun J (2013). Analyzing panel count data with a dependent observation process and a terminal event. *Canad. J. Statist* 41, 174–191.

- Zhao X, & Tong X (2011). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics & Data Analysis* 55, 291–300.
- Zhu L, Tong X, Zhao H, Sun J, Srivastava DK, Leisenring W, & Robison LL (2013). Statistical analysis of mixed recurrent event data with application to cancer survivor study. *Stat. Med* 32, 1954–1963. [PubMed: 23139023]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 1

Simulation results with independent dropout process where $T_i = C_i$.

γ		$(\beta_0, \theta_0) = (0, 0)$		$(\beta_0, \theta_0) = (0, 0.5)$		$(\beta_0, \theta_0) = (0.5, 0)$		$(\beta_0, \theta_0) = (0.5, 0.5)$	
		$\hat{\beta}_1$	$\hat{\theta}_1$	$\hat{\beta}_1$	$\hat{\theta}_1$	$\hat{\beta}_1$	$\hat{\theta}_1$	$\hat{\beta}_1$	$\hat{\theta}_1$
0	BIAS	-0.002	0.011	0.004	0.003	0	0.002	-0.014	0.001
	ESE	0.158	0.127	0.149	0.116	0.193	0.114	0.178	0.103
	SSE	0.167	0.132	0.154	0.116	0.204	0.112	0.192	0.105
	CP	0.939	0.933	0.944	0.953	0.945	0.957	0.932	0.939
0.5	BIAS	0	0	-0.010	0.004	-0.009	0.005	0	0.003
	ESE	0.152	0.123	0.141	0.112	0.185	0.110	0.172	0.101
	SSE	0.155	0.128	0.144	0.113	0.199	0.114	0.181	0.100
	CP	0.950	0.937	0.939	0.940	0.930	0.941	0.944	0.953

TABLE 2

Simulation results with dependent dropout process where $T_i = C_i \wedge D_i D_i = D_0 e^{\eta_0' W_i}$, and $D_0 = e^{\epsilon_i}$ with ϵ_i following the normal distribution $N(0.25, 0.25^2)$.

η	γ		$(\beta_0, \theta_0) = (0, 0)$		$(\beta_0, \theta_0) = (0, 0.5)$		$(\beta_0, \theta_0) = (0.5, 0)$		$(\beta_0, \theta_0) = (0.5, 0.5)$	
			$\hat{\beta}_2$	$\hat{\theta}_2$	$\hat{\beta}_2$	$\hat{\theta}_2$	$\hat{\beta}_2$	$\hat{\theta}_2$	$\hat{\beta}_2$	$\hat{\theta}_2$
0	0	BIAS	-0.006	-0.001	-0.008	-0.002	0.001	0.005	-0.015	0.004
		ESE	0.157	0.145	0.147	0.132	0.190	0.125	0.176	0.113
		SSE	0.167	0.138	0.149	0.126	0.204	0.123	0.186	0.111
		CP	0.933	0.948	0.949	0.952	0.938	0.956	0.929	0.955
0	0.5	BIAS	0.000	0.009	-0.009	0.001	-0.014	0.004	-0.010	0.004
		ESE	0.150	0.139	0.141	0.127	0.185	0.121	0.174	0.110
		SSE	0.154	0.135	0.147	0.120	0.195	0.115	0.174	0.112
		CP	0.939	0.951	0.939	0.965	0.935	0.957	0.931	0.937
0.2	0	BIAS	-0.007	-0.002	-0.003	-0.002	-0.005	-0.011	-0.014	-0.003
		ESE	0.156	0.163	0.146	0.146	0.189	0.136	0.179	0.123
		SSE	0.159	0.146	0.157	0.134	0.195	0.127	0.192	0.114
		CP	0.943	0.964	0.927	0.964	0.940	0.949	0.924	0.962
0.2	0.5	BIAS	-0.001	0.001	0.001	-0.005	-0.007	0.000	-0.010	0.007
		ESE	0.147	0.154	0.141	0.138	0.184	0.128	0.172	0.116
		SSE	0.150	0.143	0.149	0.127	0.190	0.123	0.183	0.110
		CP	0.946	0.952	0.933	0.962	0.939	0.954	0.925	0.956

TABLE 3

Empirical sizes and powers of the proposed goodness-of-fit test based on the simulated data with the independent (Scenario 1) or dependent (Scenario 2) observation process.

		$k = 0$	$k = 0.5$	$k = 1$	$k = 1.5$	$k = 2$	$k = 2.5$
Scenario 1	$n = 400$	0.039	0.249	0.690	0.931	0.989	0.996
	$n = 800$	0.053	0.521	0.956	1	1	1
Scenario 2	$n = 400$	0.041	0.200	0.526	0.805	0.940	0.980
	$n = 800$	0.042	0.368	0.865	0.986	1	1

TABLE 4

The p -values of goodness-of-fit for fitting four candidate models to the CCSS data.

Models	Test statistics		
	$\mathcal{F}_1(t; \alpha)$	$\mathcal{F}_2(t; \gamma, \eta)$	$\mathcal{F}_3(t; \eta)$
AM1+Model2+Model3	0.573	0.293	0.352
AM2+Model2+Model3	0.332	0.290	0.325
MM+Model2+Model3	0.477	0.281	0.343
AA+Model2+Model3	0.321	0.293	0.341

TABLE 5

Analysis results of the CCSS based on the MM model

		Hospital visits		Observations		Dropouts (Death)	
		INS	LOW	INS	LOW	INS	LOW
Method in Section 4	ECE	0.520	1.252	-0.119	0.525	-0.023	-0.270
	ESD	1.255	1.616	0.115	0.266	0.047	0.040
	<i>p</i> -value	0.679	0.438	0.301	0.048	0.625	0.000
Method in Section 3	ECE	0.266	0.503	-0.148	0.136		
	ESD	0.167	0.148	0.042	0.038		
	<i>p</i> -value	0.111	0.001	0.000	0.000		
Method in Hu et al.(2003)	ECE	0.339	0.521	-0.148	0.136		
	ESD	0.177	0.158	0.042	0.038		
	<i>p</i> -value	0.055	0.001	0.000	0.000		

Note: ECE = Estimated covariate effect, ESD = Estimated standard error.