Published in final edited form as:

Biometrics. 2011 September ; 67(3): 711–718. doi:10.1111/j.1541-0420.2011.01559.x.

Variance Estimation for Statistics Computed from Single Recurrent Event Processes

Yongtao Guan^{1,*}, Jun Yan^{2,3}, and Rajita Sinha⁴

¹Division of Biostatistics, Yale University, CT 06520, USA

²Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

³Institute for Public Health Research, University of Connecticut Health Center, East Hartford, CT 06108, USA

⁴Department of Psychiatry and Child Study Center, Yale University, New Haven, CT 06511, USA

SUMMARY

This paper is concerned with variance estimation for statistics that are computed from single recurrent event processes. Such statistics are important in diagnosis for each individual recurrent event process. The proposed method only assumes a semiparametric form for the first-order structure of the processes but not for the second-order (i.e., dependence) structure. The new variance estimator is shown to be consistent for the target parameter under very mild conditions. The estimator can be used in many applications in semiparametric rate regression analysis of recurrent event data such as outlier detection, residual diagnosis as well as robust regression. A simulation study and application to two real data examples are used to demonstrate the use of the proposed method.

Keywords

Inhomogeneous Point Process; Recurrent Event Process; Robust Regression; Variance Estimation

1. Introduction

Recurrent event data have become increasingly available in many fields such as biomedical studies, economics, engineering, and epidemiology (e.g., Nelson, 2003). A typical recurrent event dataset consists of realizations from many recurrent event processes, each of which may contain multiple events observed during a given study period. By definition, a single recurrent event process is a (temporal) point process.

Many existing work in recurrent event literature are concerned with statistics that are computed from the whole set of recurrent event processes; for instance, regression coefficient estimators in the widely used semiparametric models for intensity (Andersen and Gill, 1982) or for rate (Lawless and Nadeau, 1995) fall into that category. However, little has been done for statistics computed from a single recurrent event process such as

*yongtao.guan@yale.edu.

Supplementary Materials

Theoretical results and a simulation study, referenced in Section 1, are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

$$T_i(B) = \sum_{t \in (N_i \cap B)} f_i(t), \quad (1)$$

where B is a preselected time period of interest, N_i is the i th recurrent event process, i.e., a set of random event times for the i th individual, and $f_i(t)$ is a known, bounded function. For ease of presentation, we denote $f_i(\cdot)$ by $f(\cdot)$. This paper develops a consistent variance estimator for statistics in the form of (1) and discusses its applications in semiparametric regression analysis of recurrent event processes including both residual diagnostics and robust regression.

An important goal of analyzing recurrent event data is to understand and describe features in individual recurrent event processes. Statistics like (1) are characteristics of a single recurrent event process. For instance, if B is any subinterval of the followup period and $f(t) = 1$, then $T_i(B)$ is the total number of events from the i th process observed over that interval. A regression model usually provides $E\{T_i(B)\}$, i.e., the expectation of $T_i(B)$. In order to assess whether a given process is adequately described by the model, one would be interested in investigating the residual $T_i(B) - E\{T_i(B)\}$, for which an estimator for the variance of $T_i(B)$ is necessary. Standardized residuals can be constructed based on the variance estimator and be subsequently used in residual diagnostics for outlier detection and assessment of goodness of fit of a fitted model. Moreover, robust regression can be conducted by downweighting information from processes with extreme standardized residuals.

Although variance estimation for statistics like $T_i(B)$ in (1) has not been well studied for recurrent event data, several methods have been developed to estimate the variance of similar statistics derived from spatial point process data, a type of data that typically consist of only one single realization of the process (e.g., Guan, 2009a,b). These methods, however, cannot be used for recurrent event data, because they require a large number of events to be observed from each process which is often not the case for recurrent event data. For example, for the chronic granulomatous disease example in Section 5, the total number of events is 76 for a set of 128 recurrent event processes. It is thus critical to borrow information across processes in order to estimate the variance of statistics defined over a single recurrent event process. Furthermore, most frequently used semiparametric regression models have a nonparametric baseline function that complicates the analysis. Overcoming these difficulties, this paper develops a consistent variance estimator for statistics like $T_i(B)$ in (1).

The rest of the paper is organized as follows. Section 2 sets up the notation. Sections 3 and 4 discuss the proposed method and its applications in residual analysis and robust regression. Two real examples are analyzed in Section 5 and a discussion is given in Section 6. Theoretical results and a simulation study can be found in the Web-based Supplementary Materials.

2. Notation

Let $N(t)$ be the number of events that occur over the time interval $[0, t]$ from a recurrent event process N . Let $dN(t)$ be the increment $N\{t + dt\} - N(t-)$ over the small interval $[t, t + dt)$. The k th-order rate function of process N is

$$\lambda_{(k)}(t_1, \dots, t_k) = \lim_{dt_j \rightarrow 0; j \in \{1, \dots, k\}} \frac{E\{dN(t_1) \cdots dN(t_k)\}}{dt_1 \cdots dt_k}. \quad (2)$$

Intuitively, $\lambda_{(k)}(t_1, \dots, t_k) dt_1 \cdots dt_k$ is the approximate probability of having one event of N in each of the small intervals $[t_1, t_1 + dt_1), \dots, [t_k, t_k + dt_k)$. The first-order rate function $\lambda_{(1)}(\cdot)$ will be denoted by $\lambda(\cdot)$, and be abbreviated as the rate function from now on. It is a marginal quantity and specifies only the mean structure of the process, and hence is different from the intensity function (Andersen and Gill, 1982), which specifies the distribution of the process.

Let N_i , $i = 1, \dots, n$, denote n independent recurrent event processes. Suppose that N_i is observed between $[0, \tau_i]$, where τ_i is independent of N_i . Let $\lambda_i(\cdot)$ be the rate function of N_i . Conditional on some observed covariates, a semiparametric model for $\lambda_i(\cdot)$ takes the form

$$\lambda_i(t; \beta) = \lambda_0(t) \rho_i(t; \beta), \quad (3)$$

where $\lambda_0(t)$ is an unspecified baseline rate function, $\rho_i(t; \beta)$ is a parametric function of the covariates observed at t for the i th process and of a vector β of unknown parameters. The parameter vector β can be estimated by solving the estimating equations given in Lawless and Nadeau (1995) and Lin et al. (2000).

With the first- and second-order rate functions defined in (2), the pair correlation function (Møller and Waagepetersen, 2004) for a recurrent event process N is defined as

$$g(t_1, t_2) = \frac{\lambda_{(2)}(t_1, t_2)}{\lambda(t_1)\lambda(t_2)}.$$

The pair correlation function is a useful tool for characterizing the dependence in the point process N . If N is a Poisson process, then $g(t_1, t_2) = 1$ for all distinct t_1 and t_2 . Typically, $g(t_1, t_2) > 1$ indicates attraction (i.e., positive dependence) and $g(t_1, t_2) < 1$ suggests inhibition (i.e., negative dependence).

Let $g_i(t_1, t_2)$ be the pair correlation function for N_i . Assume that for some function $g(\cdot)$,

$$g_i(t_1, t_2) = g(t_1, t_2). \quad (4)$$

That is, all individual recurrent event processes share the same pair correlation function. However, the pair correlation function does not need to be stationary, as being commonly required in spatial point process literature (e.g. Møller and Waagepetersen, 2004). To see an example, suppose that conditional on an observed covariate vector process $Z_i(t)$ and a latent process $\varepsilon_i(t)$, N_i is a Poisson process with intensity $\lambda_i(t) = \lambda_0(t) \exp\{Z_i^T(t)\beta + \varepsilon_i(t)\}$. The pair correlation function then is

$$g_i(t_1, t_2) = \frac{E[\exp\{\varepsilon_i(t_1) + \varepsilon_i(t_2)\}]}{E[\exp\{\varepsilon_i(t_1)\}]E[\exp\{\varepsilon_i(t_2)\}]},$$

which is the same for all i given that the latent processes $\varepsilon_i(t)$, $i = 1, \dots, n$, have the same distribution. Following this line of construction, one can see that assumption (4) is satisfied by a wide range of recurrent event processes.

3. Variance Estimation

Let β_0 denote the true value of β . The main interest of this paper is to estimate the variance of statistics like $T_l(B)$ defined in (1). Under condition (4), it is straightforward to show that

$$\text{Var}\{T_i(B)\} = \int_{B^2} \lambda_0(t_1)\lambda_0(t_2)f(t_1)f(t_2)\rho_i(t_1;\beta_0)\rho_i(t_2;\beta_0)\{g(t_1, t_2) - 1\}dt_1dt_2 + \int_B \lambda_0(t)\{f(t)\}^2\rho_i(t;\beta_0)dt. \quad (5)$$

As described earlier, an estimate of $\text{Var}\{T_i(B)\}$ allows standardization of $T_i(B)$ and, hence, leads to residual diagnosis as well as robust regression for semiparametric regression models. A method-of-moment estimator is proposed below for the variance of $T_i(B)$ without assuming a parametric structure for the pair correlation function. Such estimators have been proposed for statistics defined on all recurrent event processes such as $\sum_{i=1}^n T_i(B)$ (e.g., Lawless and Nadeau, 1995), but not for statistics defined on a single recurrent event process such as $T_i(B)$.

The proposed estimator is presented in two steps. The second step uses the first step as building block and improves the efficiency of the estimator obtained in the first step.

3.1 A Naive Estimator

To estimate (5), consider the statistic

$$S_{ij}(B;\beta_0) = \int_B \frac{f(t)\rho_i(t;\beta_0)}{\rho_j(t;\beta_0)} dN_j(t).$$

The Campbell's Theorem (e.g., Stoyan et al., 1995) states that for any non-negative function $h(\cdot)$,

$$E\left\{\sum_{t_1, \dots, t_k \in (N \cap B)}^{\neq} h(t_1, \dots, t_k)\right\} = \int_B h(t_1, \dots, t_k)\lambda_{(k)}(t_1, \dots, t_k)dt_1 \dots dt_k,$$

where $\neq$ in the above signifies summation over distinct $t_1, \dots, t_k$ and $\lambda_{(k)}$ is the k th-order rate function defined in (2). It then follows that

$$E\{S_{ij}(B;\beta_0)\} = \int_B \lambda_0(t)f(t)\rho_i(t;\beta_0)dt = E\{T_i(B)\},$$

and

$$\text{Var}\{S_{ij}(B;\beta_0)\} = \int_{B^2} \lambda_0(t_1)\lambda_0(t_2)f(t_1)f(t_2)\rho_i(t_1;\beta_0)\rho_i(t_2;\beta_0)\{g(t_1, t_2) - 1\}dt_1dt_2 + \int_B \lambda_0(t)\frac{\{f(t)\rho_i(t;\beta_0)\}^2}{\rho_j(t;\beta_0)}dt.$$

Thus, the expected value of $S_{ij}(B; \beta_0)$ is invariant with respect to j and is equal to the expected value of $T_i(B)$. Furthermore, the variance of $S_{ij}(B; \beta_0)$ has a simple linear relationship with the target variance (5). Specifically,

$$\text{Var}\{S_{ij}(B;\beta_0)\} = \text{Var}\{T_i(B)\} - \int_B \lambda_0(t)\{f(t)\}^2\rho_i(t;\beta_0)dt + \int_B \lambda_0(t)\frac{\{f(t)\rho_i(t;\beta_0)\}^2}{\rho_j(t;\beta_0)}dt. \quad (6)$$

By averaging (6) over j where $S_{ij}(B; \beta_0)$ are well defined, it can be easily seen that

$$\begin{aligned}\text{Var}\{T_i(B)\} &= \frac{1}{K_n(B)} \sum_{j=1}^n Y_j(B) \text{Var}\{S_{ij}(B; \beta_0)\} - \frac{1}{K_n(B)} \sum_{j=1}^n Y_j(B) \int_B \lambda_0(t) \frac{\{f(t)\rho_i(t; \beta_0)\}^2}{\rho_j(t; \beta_0)} dt + \int_B \lambda_0(t) \{f(t)\}^2 \rho_i(t; \beta_0) dt \\ &\equiv V_{i1,n}(B; \beta_0) - V_{i2,n}(B; \beta_0) + V_{i3,n}(B; \beta_0),\end{aligned}$$

where $K_n(B) = \sum_{j=1}^n Y_j(B)$, $Y_j(B) = I(B \subseteq [0, \tau_j])$, and $I(\cdot)$ is an indicator function. To estimate $\text{Var}\{T_i(B)\}$, it is sufficient to estimate $V_{i1,n}(B; \beta_0)$, $V_{i2,n}(B; \beta_0)$ and $V_{i3,n}(B; \beta_0)$.

Let $\bar{S}_{i,n}(B; \beta_0) = \frac{1}{K_n(B)} \sum_{j=1}^n Y_j(B) S_{ij}(B; \beta_0)$ be an estimator for $E\{T_i(B)\}$. Note that $\text{Var}\{S_{ij}(B; \beta_0)\}$ can be estimated approximately unbiasedly, although not consistently, by the statistic $\{S_{ij}(B; \beta_0) - \bar{S}_{i,n}(B; \beta_0)\}^2$. Therefore, the estimator

$$\widehat{V}_{i1,n}(B; \beta_0) = \frac{1}{K_n(B)} \sum_{j=1}^n Y_j(B) \{S_{ij}(B; \beta_0) - \bar{S}_{i,n}(B; \beta_0)\}^2$$

is approximately unbiased for $V_{i1,n}(B; \beta_0)$. Because $\widehat{V}_{i1,n}(B; \beta_0)$ is an average of all available $\{S_{ij}(B; \beta_0) - \bar{S}_{i,n}(B; \beta_0)\}^2$, it is also consistent for $V_{i1,n}(B; \beta_0)$ under mild conditions.

To estimate $V_{i2,n}(B; \beta_0)$ and $V_{i3,n}(B; \beta_0)$, note that for any real function $h(t)$, the statistic $\int_B h(t) \rho_i(t; \beta_0) / \rho_j(t; \beta_0) dN_j(t)$ is an unbiased estimator for $\int_B \lambda_0(t) \rho_i(t; \beta_0) h(t) dt$. This fact immediately suggests that

$$\begin{aligned}\widehat{V}_{i2,n}(B; \beta_0) &= \frac{1}{K_n(B)} \sum_{j=1}^n Y_j(B) \int_B \frac{\{f(t)\rho_i(t; \beta_0)\}^2}{\{\rho_j(t; \beta_0)\}^2} dN_j(t), \\ \widehat{V}_{i3,n}(B; \beta_0) &= \frac{1}{K_n(B)} \sum_{j=1}^n Y_j(B) \int_B \frac{\{f(t)\}^2 \rho_i(t; \beta_0)}{\rho_j(t; \beta_0)} dN_j(t),\end{aligned}$$

are unbiased estimators for $V_{i2,n}(B; \beta_0)$ and $V_{i3,n}(B; \beta_0)$, respectively. As in the case of $\widehat{V}_{i1,n}(B; \beta_0)$, both estimators are consistent for their respective target parameters under mild conditions. The target variance $\text{Var}\{T_i(B)\}$ is then estimated by

$$\widehat{\text{Var}}\{T_i(B)\} = \widehat{V}_{i1,n}(B; \beta_0) - \widehat{V}_{i2,n}(B; \beta_0) + \widehat{V}_{i3,n}(B; \beta_0).$$

In practice, the value of β_0 is unknown but can be replaced by a consistent estimator $\hat{\beta}_n$, where $\hat{\beta}_n$, for instance, is obtained from an estimating equation approach.

3.2 An Improved Estimator

The naive estimator given in Section 3.1 uses only the processes satisfying $Y_j(B) = 1$, i.e., whose observation intervals include B as a subset. For the processes with $Y_j(B) = 0$, they may nevertheless still be useful for estimating the variance of $T_i(B \cap [0, \tau_j])$, which contributes partly to the variance of $T_i(B)$. A more efficient estimator should incorporate information from all processes. To that end, let B_k , $k = 1, \dots, m$, be a partition of B , i.e., B_k 's are disjoint and $B = \bigcup_{k=1}^m B_k$. For example, if $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_n$, one can then set $B_k = B \cap (\tau_{k-1}, \tau_k]$ for $k \leq i$. In this case, $m = i$. Then

$$\text{Var}\{T_i(B)\} = \sum_{k=1}^m \text{Var}\{T_i(B_k)\} + 2 \sum_{k=1}^{m-1} \sum_{l=k+1}^m \text{Cov}\{T_i(B_k), T_i(B_l)\}. \quad (7)$$

To estimate $\text{Var}\{T_i(B_k)\}$, the naive estimator can be applied directly. To estimate the covariance terms in (7), we apply Campbell's Theorem again and obtain

$$\text{Cov}\{T_i(B_k), T_i(B_l)\} = \int_{B_k} \int_{B_l} \lambda_0(t_1) \lambda_0(t_2) f(t_1) f(t_2) \rho_i(t_1; \beta_0) \rho_i(t_2; \beta_0) \{g(t_1, t_2) - 1\} dt_1 dt_2.$$

Let $B_{kl} = B_k \cup B_l$. Similar to the estimation of $V_{i1,n}(B; \beta_0)$, the above covariance can be estimated by

$$\widehat{C}_{ikl,n}(B; \beta_0) = \frac{1}{K_n(B_{kl})} \sum_{j=1}^n Y_j(B_{kl}) \{S_{ij}(B_k; \beta_0) - \bar{S}_{i,n}(B_k; \beta_0)\} \{S_{ij}(B_l; \beta_0) - \bar{S}_{i,n}(B_l; \beta_0)\},$$

where $K_n(\cdot)$ and $Y_j(\cdot)$ are as being defined in Section 3.1. In practice, β_0 is replaced by a consistent estimator β_n . Note that all possible processes with $Y_j(B_k) = 1$ are used to estimate $\text{Var}\{T_i(B_k)\}$ and all possible processes with $Y_j(B_{kl}) = 1$ are used to estimate $\text{Cov}\{T_i(B_k), T_i(B_l)\}$. Hence, the new estimator uses all available information and is therefore expected to be more efficient than the naive estimator given in Section 3.1.

4. Applications of the Proposed Method

Applications of the proposed method are discussed in the context of semiparametric modeling of the rate function for recurrent event processes, where $\lambda_i(t; \beta) = \lambda_0(t) \rho_i(t; \beta)$ as given in (3). In what follows, assume that a consistent estimator for β exists and denote it by β_n .

4.1 Outlier detection

Suppose that $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_n$ and define $B_k = B \cap (\tau_{k-1}, \tau_k]$ for $k \leq i$. Then, a fitted value of $T_i(B)$ in (1) can be obtained as

$$\widehat{T}_i(B) = \sum_{k=1}^i \frac{1}{K_n(B_k)} \sum_{j=1}^n Y_j(B_k) \int_{B_k} f(t) \frac{\rho_i(t; \widehat{\beta}_n)}{\rho_j(t; \widehat{\beta}_n)} dN_j(t).$$

Note that $\widehat{T}_i(B)$ does not involve an estimate for the nonparametric baseline rate. The residual for the i th process is defined as $R_i = T_i - \widehat{T}_i$, where T_i and $\widehat{T}_i$ denote $T_i(B)$ and $\widehat{T}_i(B)$ with $B = [0, \tau_i]$ and $f(t) = 1$. The variance estimator proposed in Section 3.2 can be used to obtain a standardized residual $r_i = R_i / \{\widehat{\text{Var}}(T_i)\}^{1/2}$, assuming that the variance of $\widehat{T}_i$ is small. An extremely large value of r_i signifies the i th process being an outlier.

In a standard regression setting, an observation is often called an outlier if the absolute value of its standardized residual is larger than 2 or 3. However, such a practice may not be valid in the current setting, because the distribution of T_i may be far from being normal. To derive a more proper decision rule, assume as a compromise that the i th recurrent event process is Poisson with mean $\widehat{T}_i$. Let $Q_i(\alpha)$ be the $100(1 - \alpha)$ th percentile from the resulting Poisson

distribution. The i th process is called an outlier if $r_i > r_{\text{cutoff}}$, where $r_{\text{cutoff}} = \{Q_i(\alpha) - \widehat{T}_i\} / \sqrt{\widehat{T}_i}$. Note that this implies that $T_i - \widehat{T}_i > r_{\text{cutoff}} \sqrt{\widehat{\text{Var}}(T_i)}$.

In the case that the i th recurrent event process is clustered, the actual $100(1 - \alpha)$ th percentile of the process is typically larger than its counterpart from a Poisson process with the same mean, i.e., $Q_i(\alpha)$. Because $\widehat{\text{Var}}(T_i)$ also tends to be larger than $\widehat{T}_i$, the proposed decision rule at least partially accounts for the effect due to clustering by comparing $(T_i - \widehat{T}_i)$ against a value that is larger than $\{Q_i(\alpha) - \widehat{T}_i\}$. Thus, the Poisson assumption is not overly restrictive. The method can also be used to define residuals over any subset of $[0, \tau_i]$, e.g., over a moving time window within $[0, \tau_i]$. Such an analysis can provide additional information about whether a given recurrent event process is extreme only in a subset of the followup period or throughout the whole period; see Section 5 for illustrations.

4.2 Assessment of goodness of fit

In a standard regression analysis, plots of standardized residuals are routinely evaluated to assess the adequacy of a fitted model. The residual plots proposed in Section 4.1 are not suitable for this purpose because of the highly non-normal nature of the residuals. We instead propose a “binned” residual plot approach by combining residuals based on proximity in their fitted/predictor values, so that the combined residuals are closer to being normal. For illustration, consider below the case that $\rho_i(t; \beta) = \rho_i(\beta)$, i.e., the parametric part of the rate function does not vary with time.

Let X_i be a generic variable which could be either the fitted value or a predictor variable or an omitted variable, all of which are associated with the i th process. Assume that $X_1 \leq X_2 \leq \dots \leq X_n$. We first start from X_1 and stop at the smallest index I_1 such that $\sum_{i=1}^{I_1} \widehat{T}_i \geq k$ for some predefined k , i.e., the estimated expected number of events from the first I_1 processes is larger than or equal to k . We next start from X_{I_1+1} and stop at the smallest index I_2 such that $\sum_{i=I_1+1}^{I_2} \widehat{T}_i \geq k$, and so on and so forth. Write $I_0 = 0$. The j th “binned” standardized residual is defined as

$$r_{(j)} = \frac{\sum_{i=I_{j-1}+1}^{I_j} (T_i - \widehat{T}_i)}{\left\{ \sum_{i=I_{j-1}+1}^{I_j} \widehat{\text{Var}}(T_i) \right\}^{1/2}},$$

where the variances in the denominator are obtained by the proposed variance estimator in Section 3.2. Also define

$$\bar{X}_{(j)} = \frac{1}{I_j - I_{j-1}} \sum_{i=I_{j-1}+1}^{I_j} X_i.$$

The “binned” standardized residuals $r_{(j)}$ can be plotted against $\bar{X}_{(j)}$ for residual diagnostics. If k is reasonably large but not too large, say $k = 30$ together with $f(t) = 1$, then $r_{(j)}$ is approximately normally distributed and $\bar{X}_{(j)} \approx X_i$, $i = I_{j-1} + 1, \dots, I_j$. These two facts combined suggest that residual analysis can be conducted as usual to assess the goodness of fit of a fitted model. However, $k = 30$ entails that the total number of events should be large. Thus, the “binned” standardized residuals may not be used for data with a small sample size,

such as the chronic granulomatous disease data given in Section 5.1. For a given problem, one may examine the distribution of the bins so as to assess whether the current sample size is large enough to reveal useful information regarding the goodness of fit.

4.3 Robust estimation of regression parameters

In a standard regression analysis, weighted least squares estimation is often used to reduce the influence of outliers and hence to obtain more robust estimates of the regression parameters. The idea is to assign different weights to the observations based on their standardized residual values. A popular weight function is the Huber weight function: $w(u) = 1$ if $|u| \leq 1.345$ and $w(u) = 1.345/|u|$ otherwise. When conducting the analysis, $w(u)$ defines the weight applied to an observation with a standardized residual u .

For recurrent event analysis, we follow Binder (1992) to solve

$$U(\beta) = \sum_{i=1}^n w_i \int_0^{r_i} \frac{\rho_i^{(1)}(t; \beta)}{\rho_i(t; \beta)} dN_i(t) - \sum_{i=1}^n w_i \int_0^{r_i} \frac{\sum_{j=1}^n w_j Y_j(t) \rho_j^{(1)}(t; \beta)}{\sum_{j=1}^n w_j Y_j(t) \rho_j(t; \beta)} dN_i(t) = 0. \quad (8)$$

Note that if the weights, $w_i : i = 1, \dots, n$, are constant, then (8) simply reduces back to the estimating equations used in Lin et al. (2000). For fixed weights, (8) forms a set of unbiased estimating equations and hence solving it can yield a consistent estimator for β . Because observations with smaller weights have less effect on the value of the estimating equations, (8) can be used to downweight the influence of outlying cases by assigning them relatively small weights.

Although the Huber weight function can be used to obtain the weights, it is designed to be most efficient for normal data. Given the non-normal nature of the recurrent even data, some modifications of the standardized residuals are necessary before they can be used to derive the weights. One possibility is to first normalize them, e.g., by using $v_i = (r_i - \min_j r_j + \theta)^\eta$ for some $\theta > 0$ and $\eta \neq 0$, $i = 1, \dots, n$, and then define

$$u_i = (v_i - \bar{v}) / s_v, \quad (9)$$

where $\bar{v}$ and s_v are the sample mean and sample standard deviation of v_i . Alternatively, one can directly standardize T_i by using the delta method with a squared root transformation, which leads to

$$u_i = \frac{T_i^{1/2} - \widehat{T}_i^{1/2}}{\{\widehat{\text{Var}}(T_i) / (4\widehat{T}_i)\}^{1/2}}. \quad (10)$$

To enhance robustness, both u_i 's can be obtained by leaving the i th process out. The Huber weight function can then be used to obtain the weights based on u_i , $i = 1, \dots, n$. We evaluate the performance of these two proposals through simulation. To make inferences about the robust estimator, we will rely on standard errors from bootstrap.

5. Real Examples

5.1 Chronic Granulomatous Disease Data

Chronic granulomatous disease is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections. The data came from a placebo controlled

trial of gamma interferon in chronic granulomatous disease. It consists of 128 patients from 13 hospital centers. The semiparametric marginal rate model (3) is used with the parametric part of the function being given by $\rho_i(t) = \exp\{Z_i(t)^T \beta\}$, where $Z_i(t)$ is a covariate vector. Two covariates, treatment indicator and patient age, are considered as in Lin et al. (2000).

Figure 1 shows the standardized residuals against the fitted values at the individual process level and at the center level. Also plotted are the lines of monitoring bounds

$\{Q(\alpha) - \widehat{T}_i\} / \sqrt{\widehat{T}_i}$ with $\alpha = 0.05$ (lower) and 0.01 (upper). These lines are zigzag because of the discreteness of Poisson variables. At the individual patient level, there are seven standardized residuals exceeding the 0.05 bound; see Figure 1, panel (a). The most extreme standard residual comes from the subject with id 36, a thirteen-year old individual from the treatment group who had two events in 238 followup days. The subject would not have been spotted suspicious if one looked only for a large number of event counts. A similar standardized residual plot is extended to the center level; see Figure 1, panel (b). None of the centers has standardized residual above 2, with the majority between -1 and 1 . The two largest standard residuals come from center id 204 (Scripps Institute, with sixteen patients) and center id 243 (Mott Children's Hospital, with nine patients). Standardized residual from Center 204 is barely outside of the 0.05 bound. Absence of obvious patterns and extreme values suggests that there is no need to include a random effect term for center effect. Since the total number of events is quite low in this dataset, a "binned" residual plot is of little use here.

Robust estimates of the regression coefficients as well as those from the usual Cox regression are summarized in the upper panel of Table 1. The robust estimates of the treatment effect are slightly bigger in magnitude than that from the usual Cox regression. Their standard errors are obtained from 100 bootstrap samples. Although the standard errors of the robust estimates are higher, the treatment effect is still significant. The age effect is marginally significant in the usual Cox regression, but is no longer so from either robust estimates.

5.2 Cocaine Relapse Data

Cocaine dependence remains a serious public health problem in the US with more than one million individuals meeting criteria for current dependence (SAMHSA, 2004). The cocaine relapse data being considered here consist of daily relapse patterns for 90 days from 110 subjects, who were prospectively followed after discharge from an inpatient treatment for cocaine dependence. Four pretreatment variables are used as covariates, which include baseline cocaine use frequency (bs.freq) over a 90-day baseline period, cocaine use years (bs.years), gender (1 for female and 0 for male) and race (1 for Caucasian and 0 otherwise). To capture any potential nonlinear effect of baseline use frequency, a squared term (bs.freq²) is also included in the model.

The first row of Figure 2 shows the standardized residuals. Panel (a) plots the standardized residuals against fitted values for all individual subjects along with two monitoring bounds for $\alpha = 0.05$ and 0.01. As the fitted values have a wide range up to around 45, the monitoring bounds are observed to converge to the respective quantiles of a standard normal distribution. There are four subjects with standardized residuals exceeding the 0.05 bound, two of which also exceed the 0.01 bound. The most extreme standardized residual is around 7, from the subject with id 17, a non-Caucasian male whose baseline frequency is about the 6th percentile among all the subjects, but whose relapse count is 34 in 67 days of followup, much higher than expected. This subject is very likely an outlier and the cause for it needs further investigation. Panels (b) plots the standardized residuals for three suspected outliers, subjects 17, 32, and 25, and over moving windows $B = [t - 10, t + 10] \cap [0, \tau_i]$, $0 < t \leq \tau_i$. Subject 6 is also a potential outlier but it is not plotted because the followup lasted only 13

days. The plot reveals quite different cocaine use patterns among the three subjects. Specifically, subject 17 started with a much higher than normal cocaine use frequency from the beginning and stayed that way throughout the followup period; subject 25 started with a normal frequency which then later gradually increased a modestly higher than normal frequency roughly after day 20; in great contrast, subject 32 began with a very low cocaine use frequency until around day 40 and then drastically accelerated to an extremely high frequency after that. These findings are interesting and practically meaningful but would not have been made by using the plot based on $B = [0, \tau_i]$ alone.

The second row of Figure 2 plots the “binned” standardized residuals against the fitted values and the covariate baseline cocaine use frequency, with threshold value $k = 30$ and $f(\cdot) = 1$. These plots resemble residual plots commonly used in a standard regression analysis and can in fact be interpreted in the same way. Neither plot suggests any evidence for lack of fit. Further residual diagnosis has been performed by examining plots of “binned” standardized residuals against the remaining covariates in the model (not shown). No obvious pattern is found from any of these plots, suggesting that the model provides a good fit.

Robust estimates of the regression coefficients are summarized in the lower panel of Table 1. Among the available predictors, baseline use frequency is significant at the 0.01 level and race is significant at the 0.1 level in both robust regressions and the usual Cox regression, whereas gender is significant at the 0.1 level in the robust regressions but not in the Cox regression. Because the signs for all three regression coefficients are positive, the results suggest that women, Caucasians and more frequent baseline users tend to relapse more than men, non-Caucasians and less frequent baseline users. The finding that women relapse more is likely because they often face significant social stressors during recovery, which may increase their risk of relapse after treatment (Miller et al., 1999). The finding that Caucasians relapse more is surprising, because it has been reported, although under a different study design, that African Americans are less likely to be abstinent at discharge, a sign for more severe future relapse (Dodge et al., 2005). The cause for the conflicting findings needs to be further investigated. The conclusion that more frequent baseline users relapse more agrees with the recent finding that high-frequency cocaine users show significantly higher levels of craving and anxiety in drug-cue conditions, which expectedly will cause higher relapse rates, as compared with low-frequency users (Fox et al., 2005). Note that the magnitudes of the robust estimates are larger than the estimates in the usual Cox regression, indicating a stronger effect on relapse from these variables. The difference is at least partly attributed to subject 17 and other extreme subjects, and assigning lower weights to them is intuitively appealing. The negative but insignificant estimated coefficient of the squared baseline use frequency implies that the slope of baseline use frequency is decreasing and in fact flattens at higher values. The squared term is kept in the final model in order to achieve a better fit.

6. Discussion

This paper provides consistent variance estimators for statistics arising from single recurrent event processes. Although nonparametric variance estimation has been studied in literature for statistics computed from all recurrent event processes, little has been done to estimate variance of statistics defined on a single process such as (1). The proposed variance estimation method can be applied in many settings where knowledge on variance of statistics from a single recurrent event process may be important, for example, residual diagnostics and robust rate regression as being discussed in Section 4.

Lin et al. (2000) presented a set of residual diagnostic tools for semiparametric mean/rate models for recurrent event data. The statistic processes used in their analysis, when

evaluated at any fixed point, are sums of many individual residuals. Our proposed residual analysis approach is different from theirs, since we use individual standardized residuals that are derived from our proposed variance estimator for individual residuals. Similar to Lin et al.'s approach, our approach can be used to assess the validity of the functional form of covariates. Different from theirs, it can also be used for outlier detection, but not for checking the correctness of the link function or the proportionality assumption. In view of these facts, we believe that our approach is a useful supplementary to Lin et al.'s.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NSF grant DMS-0845368 and by NIH grants 1R01DA029081 and UL1DE019586. The authors thank the reviewers for their helpful comments and suggestions.

References

- Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study (Com: P1121–1124). *The Annals of Statistics*. 1982; 10:1100–1120.
- Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika*. 1992; 79:139–147.
- Dodge R, Sindelar J, Sinha R. The role of depression symptoms in predicting drug abstinence in outpatient substance abuse treatment. *Journal of Substance Abuse Treatment*. 2005; 28:189–196. [PubMed: 15780549]
- Fox HC, Talihb M, Malisona R, Anderson GM, Kreek MJ, Sinha R. Frequency of recent cocaine and alcohol use affects drug craving and associated responses to stress and drug-related cues. *Psychoneuroendocrinology*. 2005; 30:880–891. [PubMed: 15975729]
- Guan Y. Fast block variance estimation procedures for inhomogeneous spatial point processes. *Biometrika*. 2009a; 96(1):213–220.
- Guan Y. On nonparametric variance estimation for second-order statistics of inhomogeneous spatial point processes with a known parametric intensity form. *Journal of the American Statistical Association*. 2009b; 104(488):1482–1491. [PubMed: 20336182]
- Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics*. 1995; 37:158–168.
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B, Methodological*. 2000; 62(4): 711–730.
- Miller BA, Smyth NJ, Mudar PJ. Mother's alcohol and other drug problems and their punitiveness toward their children. *Journal of Studies on Alcohol*. 1999; 60(5):632–642. [PubMed: 10487732]
- Møller, J.; Waagepetersen, RP. *Statistical Inference and Simulation for Spatial Point Processes*. New York: Chapman & Hall; 2004.
- Nelson, WB. *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. Philadelphia: Society for Industrial and Applied Mathematics; 2003.
- SAMHSA. National household survey on drug abuse. 2004. <http://www.oas.samhsa.gov/nhsda.htm>.
- Stoyan, D.; Kendall, WS.; Mecke, J. *Stochastic Geometry and Its Applications*. Chichester: John Wiley & Sons; 1995.

Figure 1. Residual plots for chronic granulomatous disease data. The zigzag lines in each plot are the 95% (lower) and 99% (upper) confidence bands, respectively.

Figure 2.

Residual plots for the cocaine relapse data. Specifically, (a) plots standardized residuals against fitted values, (b) plots standardized residuals against time for three suspected outliers with moving windows $B = [t - 10, t + 10] \cap [0, \tau_i]$, $0 < t \leq \tau_i$, and (c) and (d) plot binned residuals against fitted values and baseline cocaine use frequency, respectively. In plot (a), the zigzag lines are the 95% (lower) and 99% (upper) confidence bands, respectively. In plot (b), points exceeding the 0.95 boundary are in circle, and points exceeding the 0.99 boundary are in solid circles.

Table 1

Robust and the usual Cox regression estimates (ESTs) of regression coefficients for the two real examples. The standard errors (SEs) are obtained from 100 bootstrap samples.

Variable	SQRT		DLT		Cox	
	EST	SE	EST	SE	EST	SE
<i>Chronic Granulomatous Disease Data</i>						
treatment	-1.335	0.470	-1.418	0.543	-1.149	0.349
age	-0.035	0.019	-0.036	0.021	-0.030	0.015
<i>Cocaine Relapse Data</i>						
gender	0.670	0.377	0.718	0.398	0.506	0.314
race	0.720	0.422	0.753	0.442	0.593	0.348
bs.years	-0.046	0.192	-0.036	0.202	-0.071	0.160
bs.freq	0.691	0.208	0.736	0.212	0.588	0.186
bs.freq ²	-0.353	0.253	-0.387	0.262	-0.267	0.213