A novel approach for analyzing data on recurrent events with duration to estimate the combined cumulative rate of both variables over time

Sudipta Bhattacharya

Biometrics and Information Sciences, 101 ORD, One MedImmune Way, Gaithersburg, 20878, United States

ARTICLE INFO

ABSTRACT

Recurrent adverse events, once occur often continue for some duration of time in clinical trials; and the number of events along with their durations is clinically considered as a measure of severity of a disease under study. While there are methods available for analyzing recurrent events or durations or for analyzing both side by side, no effort has been made so far to combine them and present as a single measure. However, this single-valued combined measure may help clinicians assess the wholesome effect of recurrence of incident comprising events and durations. Non-parametric approach is adapted here to develop an estimator for estimating the combined rate of both, the recurrence of events as well as the event-continuation, that is the duration per event. The proposed estimator produces a single numerical value, the interpretation and meaningfulness of which are discussed through the analysis of a real-life clinical dataset. The algebraic expression of variance is derived, asymptotic normality of the estimator is noted, and demonstration is provided on how the estimator can be used in the setup of testing of statistical hypothesis. Further possible development of the estimator is also noted, to adjust for the dependence of event occurrences on the history of the process generating recurrent events through covariates and for the case of dependent censoring.

1. Introduction

In clinical trials on diseases like Chronic Obstructive Pulmonary Disease (COPD), asthma, or migraine, etc., the event-durations are of interest along with the event-counts, as together they define severity of the disease.

Poisson regression or Negative-Binomial regression described by Lawless [1] for analyzing data on recurrent events when covariates are considered not time dependent; or for time dependent covariates, estimating the mean or rate function of recurrent events, *e.g.*, method introduced by Lin et al. [2,3] and by Miloslavsky et al. [4] (all the three based on the definition of intensity function introduced by Andersen-Gill [5]) are the standard approaches. Otherwise, if event occurrence is considered dependent on previous events, then stratified Andersen-Gill model (Cook and Lawless [6], pp 175–176) can be used. In addition, non-parametric Nelson Aalen estimator ([7]) for the rate or mean function of recurrent events, and the extensions by Cook et al. [8]) for event dependent censoring and termination are commonly used methods for analyzing data on recurrent events.

For the analysis of waiting times (with assumption of independence among waiting times and deviating from that assumption), detailed discussion is provided in chapter 4 of Cook and Lawless [6], pp 121–160. Otherwise, the modeling of proportional hazard ratio using

stratified Cox-type models ([9]) based on total time as well as on gap times introduced by Prentice, William and Peterson [10] and marginal Cox-models based on total time introduced by Wei, Lin and Weisfeld [11] are used as well.

On methods for analyzing data on duration, Metcalfe et al. [12] made a thorough coverage in their article. Otherwise, X. Joan Hu et al. [13] also proposed some methods for analyzing event-duration. The bivariate approach to deal with recurrent events with duration is through an alternating two-state process ('exacerbation state' and 'exacerbation-free state' being the two alternative states) as described by Cook and Lawless ([6] section 6.5, pp 216–218 and section 6.7.2, pp 232–236).

However, none of the methods mentioned above present an estimate for combined cumulative rate or mean of recurrent events and duration of events over time.

Here in this paper, a non-parametric estimator is proposed that takes the totality of the data into account through dealing with both, the recurrence of events and the duration of them simultaneously; and as a result, produces a single numerical value, which estimates the wholesome effect of the incident. Consequently, the proposed estimator can be looked upon as a joint or combined rate of both, the event recurrences as well as the duration per event.

Following is how the concept of the proposed estimator is developed

over the next few sections. In section 2, the mathematical motivation, development and properties of the estimator are described. In section 3, the interpretation, usefulness and meaningfulness of the single value, produced by the estimator based on a real-life clinical dataset are discussed; and in section 4, the possible use and advantages of the proposed estimator are described and further potential developments are mentioned.

2. Mathematical development

The mathematical framework for this paper is built on a process that generates recurrence of events to individual subjects, who constitute a population; and the assumption regarding occurrence of events here is that, an event can occur and end both at a time instant t (like as it happens in case of any Poisson process), or can occur at one time instant and then continues for some time before it ends.

2.1. Definitions

Let us denote by X_t the number of events that are observed to have occurred to an individual subject *by* (*i.e.*, on or before) time t.

The intensity function is defined (Cook and Lawless [6], chapter 1.3, p 10) as

$$\chi(t|\mathcal{H}_t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(\Delta X_t = 1 \mid \mathcal{H}_t)}{\Delta t} \qquad (1)$$

where $\Delta X_t = X_{t+\Delta t} - X_t$ and $\mathcal{H}_t = \{X(s), 0 \leq s < t\}$ is the history of a process.

Note that the intensity function can also be looked upon as $\chi(t \mid \mathcal{H}_t)dt = P(dX_t = 1 \mid \mathcal{H}_t) = E\{dX_t \mid \mathcal{H}_t\}$.

Based on the definition of intensity function for the occurrence (or, onset, to be precise) of a new event presented above, let us define the intensity function for ending of events for an event that has already occurred (and started) to an individual subject at time-point $\tilde{t}_0 < t$ (*i.e.*, the onset of the event was at time-point $\tilde{t}_0 < t$) and is not continuing *until* (*i.e.*, has ended *by*) time t as:

$$\chi^Z(t|\mathcal{H}_t) = \lim_{\Delta t \downarrow 0} \frac{Pr(-\Delta Z_t = 1|\mathcal{H}_t)}{\Delta t} \qquad (2)$$

where $\mathcal{H}_t = \{X(s), 0 \leq s < t\}$ is the history of a process and $\Delta Z_t = Z_{t+\Delta t} - Z_t$, with Z_t denoting an indicator function such that,

$Z_t = 1$, when an event that has already occurred (and started) at time-point $\tilde{t}_0 < t$ (*i.e.*, the onset of the event was at time-point $\tilde{t}_0 < t$), continues *until* time t, or,

$Z_t = 0$, when an event that has already occurred (and started) at time-point $\tilde{t}_0 < t$ (*i.e.*, the onset of the event was at time-point $\tilde{t}_0 < t$), has also ended *by* time t.

Note that the intensity function can also be looked upon as $\chi^Z(t|\mathcal{H}_t)dt = Pr(-\Delta Z_t = 1|\mathcal{H}_t) = E\{-dZ_t \mid \mathcal{H}_t\}$.

2.2. Mathematical motivation

Let us now define the following variables:

N_t = total count of the onset of events occurred to the population of n subjects *by* time t, which is non-decreasing over time, and
N_t^Z = count of events that have already occurred to the population of n subjects before time t and are continuing *until* time t.

Clearly, $N_t \geq N_t^Z$ at any given time t.
If we define **a new variable N_t^s** as:
$\mathbf{N_t^S = N_t + N_t^Z}$, then $\Delta N_t^S = \Delta N_t + \Delta N_t^Z$.
Then, $\Delta N_t = \Delta N(t) = \sum_{k=1}^n [X^{(k)}(t + \Delta t) - X^{(k)}(t)] Y^{(k)}(t) C^{(k)}(t)$,
where $X^{(k)}(t)$ is an indicator function such that:

$X^{(k)}(t) = 1$ when a new event has occurred (in terms of onset of that new event) *by* time t to the kth individual subject, or
$X^{(k)}(t) = 0$, when a new event has not occurred (in terms of onset of that new event) *until* time t since the preceding event occurred and ended to the kth individual subject;
$Y^{(k)}(t)$ is an indicator function with following such that:
$Y^{(k)}(t) = 1$, when the kth subject belongs to the risk set at time t for having a new event, since the preceding event occurred and ended,
$Y^{(k)}(t) = 0$, when the kth subject does not belong to the risk set at time t for having a new event, since the preceding event has occurred and is continuing;
and $C^{(k)}(t)$ is an indicator function with following such that:
$C^{(k)}(t) = I(t \leq C^{(k)})$ is an indicator function of whether the kth subject is under observation at time t.

Clearly, $C^{(k)}(t)$ is the indicator for censoring of a subject and here we assume data to be missing at random after censoring.

We also consider $\Delta N_t^Z = \sum_{k=1}^n [1 - \{Z^{(k)}(t) - Z^{(k)}(t + \Delta t)\}] Z^{(k)}(t) C^{(k)}(t)$, where $Z^{(k)}(t)$ is an indicator function such that:

$Z^{(k)}(t) = 1$ if an event that has already occurred (and started) to the kth subject and is continuing *until* time t;
$Z^{(k)}(t) = 0$ if an event that has already occurred (and started) to the kth subject and has also ended *by* time t.

It should be noted that since $Y^{(k)}(t) + Z^{(k)}(t) = 1$ at any given time t for any k,

$$\sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t) + \sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t) = \sum_{k=1}^n \{Y^{(k)}(t) + Z^{(k)}(t)\} C^{(k)}(t)$$
$$= \sum_{k=1}^n C^{(k)}(t). \qquad (3)$$

2.3. Development of the estimator

It is already defined that $N_t^S = N_t + N_t^Z$, implies $\Delta N_t^S = \Delta N_t + \Delta N_t^Z$, where $\Delta N_t = N_{t+\Delta t} - N_t$.

Since by the total probability theorem, $P(a \cup b) = P(\varepsilon) P(a \cup b \mid \varepsilon) + P(\varepsilon^C) P(a \cup b \mid \varepsilon^C) = P(\varepsilon) P(a \mid \varepsilon) + P(\varepsilon^C) P(b \mid \varepsilon^C)$, in case a and b are disjoint and $P(b \mid \varepsilon) = 0$ and $P(a \mid \varepsilon^C) = 0$, Where

ε: risk set of subjects at time t for having a new event
ε^C: set of (subjects with) existing events (*i.e.*, events that did not end *by* time t) that continuing *until* time t
a: occurrence of a new event to a subject within the interval of $[t, \ t + \Delta t)$
b: an existing event continuing during the interval of $[t, \ t + \Delta t)$

Which implies that $\frac{\Delta N_t^S}{\sum_{k=1}^n C^{(k)}(t)} = \frac{\Delta N_t + \Delta N_t^Z}{\sum_{k=1}^n C^{(k)}(t)} = \frac{\Delta N_t}{\sum_{k=1}^n C^{(k)}(t)} + \frac{\Delta N_t^Z}{\sum_{k=1}^n C^{(k)}(t)}$

$$= \left[\frac{\sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t)}{\sum_{k=1}^n C^{(k)}(t)} \right] \times \frac{\Delta N_t}{\sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t)}$$
$$+ \left[\frac{\sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t)}{\sum_{k=1}^n C^{(k)}(t)} \right] \times \frac{\Delta N_t^Z}{\sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t)},$$

where $\sum_{k=1}^n C^{(k)}(t) = \sum_{k=1}^n Y^{(k)}(t) C^{(k)}(t) + \sum_{k=1}^n Z^{(k)}(t) C^{(k)}(t)$, from equation (3).

Integrating over time, $\int_0^t \frac{dN_u^S}{\sum_{k=1}^n C^{(k)}(u)} = \int_0^t \frac{dN_u + dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)} = \int_0^t \frac{dN_u}{\sum_{k=1}^n C^{(k)}(u)} + \int_0^t \frac{dN_u^Z}{\sum_{k=1}^n C^{(k)}(u)}$