Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Semiparametric Estimation of a Change-point for Recurrent Events Data

Daniel Frobish, Nader Ebrahimi & Dung Pham

To cite this article: Daniel Frobish, Nader Ebrahimi & Dung Pham (2016) Semiparametric Estimation of a Change-point for Recurrent Events Data, Communications in Statistics - Simulation and Computation, 45:9, 3339-3349, DOI: 10.1080/03610918.2014.944654

To link to this article: http://dx.doi.org/10.1080/03610918.2014.944654

Accepted author version posted online: 12 Dec 2014.
Published online: 12 Dec 2014.

Communications in Statistics—Simulation and Computation®, 45: 3339–3349, 2016
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2014.944654

Semiparametric Estimation of a Change-point for Recurrent Events Data

DANIEL FROBISH,[1] NADER EBRAHIMI,[2] AND DUNG PHAM[1]

[1]Department of Statistics, Grand Valley State University, Allendale, New Jersey, USA
[2]Department of Statistics, Northern Illinois University, Dekalb, Illinois, USA

One particular recurrent events data scenario involves patients experiencing events according to a common intensity rate, and then a treatment may be applied. The treatment might be effective for a limited amount of time, so that the intensity rate would be expected to change abruptly when the effect of the treatment wears out. In particular, we allow models for the intensity rate, post-treatment, to be at first decreasing and then change to increasing (and vice versa). Two estimators of the location of this change are proposed.

Keywords Change-point; Recurrent events; Survival analysis.

Mathematics Subject Classification Primary 62; Secondary 62NXX

1. Introduction

Repeated events processes, where the subject experiences the same type of event more than once, are common in various applied fields such as reliability, medicine, social sciences, business, and criminology. Recurrence data consist of the times to any number of repeated events for each sample unit, for example, times to recurrent episodes of a disease in patients or times of repair of a manufactured product. The sample units are considered to be statistically independent, but the times between events within a sample unit are not necessarily independent nor identically distributed. The data are usually censored in the sense that sample units have different ends of histories. For simplicity of our presentation, in this article we focus on medical applications (e.g., seizures, heart attacks, cancerous tumors, etc.). We refer to Pena and Stocker (2007), Pena et al. (2007) and Gonzalez et al. (2010), and references cited there for more discussions of recurrent event data. Another good source for discussion is Cook and Lawless (2007).

The mathematics behind analysis of recurrence data involves the theory of counting processes. A counting process $\{N(t); t \geq 0\}$ is a nondecreasing stochastic process that has jumps of size one each time an individual experiences an event of interest. The process $N(t)$ is continuous (and constant) almost surely, except at event times. At these times, $N(t)$ is taken to be continuous from the right with a limit from the left (cadlag). Throughout this article, we assume there are no simultaneous events for a given unit. Thus, if t is an event

Received January 31, 2014; Accepted July 8, 2014
Address correspondence to Daniel Frobish, Grand Valley State University, Allendale, NJ, USA; E-mail: frobishd@gvsu.edu

time, then $N(t) - N(t-) = 1$. According to the Doob–Meyer decomposition theorem, see Anderson et al. (1993), $N(t)$ can be decomposed into a sum of two stochastic processes, $\Lambda(t)$ and $M(t)$, i.e., $N(t) = \Lambda(t) + M(t)$, where $\Lambda(t)$ is the compensator of $N(t)$ or the cumulative intensity process of $N(t)$. The theorem guarantees that $\Lambda(t)$ is a predictable, right-continuous process (but not necessarily deterministic), and $M(t)$ is a martingale. If it exists, define $\lambda(t) = \frac{d\Lambda}{dt}$ to be the intensity rate. Equivalently, one can define

$$\lambda(t) = \lim_{\delta \to 0} \frac{P(N[t, t+\delta) \geq 1 | \mathcal{F}_{t-})}{\delta}, \tag{1.1}$$

where $\mathcal{F}_{t-}$ is the complete history of the process, sometimes called the filtration, just prior to time t. Thus, given the history prior to time t, $\delta\lambda(t)$ is the approximate probability of observing at least one event in a small time interval of width δ.

Another related quantity is the expected number of events up to time t, referred to as the mean cumulative function, $E(N(t)|\mathcal{F}_{t-})$. The mean cumulative function is usually of interest in the analysis of recurrence data, just as the mean function in regression is modeled. Assuming that it exists, define $\mu(t) = dE(N(t)|\mathcal{F}_{t-})/dt$. Because we assume no two events can occur simultaneously, $\lambda(t) = \mu(t)$. That is, $\Lambda(t) = E(N(t)|\mathcal{F}_{t-})$. Thus, when we model $\Lambda(t)$, we really are modeling the mean cumulative function.

In some practical situations, $\lambda(t)$ can change abruptly, so as to cause it to be discontinuous at one or more points. Alternatively, the rate can change from decreasing to increasing or vice versa. For example, application of a treatment, environmental change, or in general, any commonly experienced event can cause such a shift. The point in time of the shift is called a change-point. See Karasoy and Kadilar (2007) and Wu et al. (2003) and references cited there for a discussion of similar change-point issues in the context of hazard rates. More recently, Zhang et al. (2014) developed procedures for change-points in hazard rates for long-term survivors. In this article, a nonparametric method is proposed to estimate such a change-point, when intensity rates change from decreasing to increasing or increasing to decreasing.

More specifically, consider m patients experiencing recurrent events according to a common intensity rate

$$\lambda(t) = \lambda_0 I_0(t) + \{\gamma_1(t-\tau_1) + \lambda_0\} I_1(t) + \{\gamma_2(t-\tau_2) + \gamma_1(\tau_2-\tau_1) + \lambda_0\} I_2(t), \tag{1.2}$$

where $I_0(t) = I(0 \leq t < \tau_1)$, $I_1(t) = I(\tau_1 \leq t < \tau_2)$ and $I_2(t) = I(\tau_2 \leq t < \infty)$, $\gamma_1(t)$ is a decreasing (increasing) function on (τ_1, τ_2) and $\gamma_2(t)$ is an increasing (decreasing) function for $t > \tau_2$, and further, $\lambda(t)$ is continuous. Without loss of generality, we assume that each patient begins the study at time 0. In Equation (1.2), τ_1 is known while τ_2 is considered unknown, in keeping with the idea that patients undergo a treatment at τ_1 (known), and then the effectiveness of the treatment may cease to exist at τ_2 (unknown). In this article, we will demonstrate two methodologies to estimate τ_2. Recently, Frobish and Ebrahimi (2007) and Ashcar et al. (2007) proposed methods to estimate τ_2 when both $\gamma_1(t)$ and $\gamma_2(t)$ are assumed to be free of t.

Remark 1.1. It should be emphasized that no parametric assumptions are made about the functions $\gamma_1(t)$ and $\gamma_2(t)$. Further, λ_0 can be made to be a function of time without causing any major changes in these methods.

In Section 2, we propose two methodologies for estimating the change-point τ_2. The first one is based on the Nelson–Aalen estimator (see Andersen et al. 1993), and the second is based on a modified Nelson–Aalen estimator. Properties of our estimates are discussed in Section 2. In Section 3, through simulations we explain how to implement our procedures. In Section 4, an example adapted from Dibley et al. (1996) is analyzed to illustrate the proposed ideas.

2. Estimation of τ_2

In this section, we propose methods to estimate τ_2. An important feature of our methodologies is that we do not make any assumptions regarding the forms of $\gamma_1(t)$ and $\gamma_2(t)$ in Equation (1.2).

To describe our methodologies, we first assume that, in Equation (1.2), $\gamma_1(t)$ and $\gamma_2(t)$ are decreasing and increasing functions on (τ_1, τ_2) and (τ_2, ∞), respectively. Then, it is clear that the true cumulative intensity function $\Lambda(t)$ is concave before τ_2 and convex after τ_2.

Throughout this article, we assume that a sample consists of arrival times, t_{ij}, $i = 1, \ldots, n_j$ for the jth patient, $j = 1, \ldots, m$, as well as the end of observation times, C_j. We also combine all the arrival times to order the distinct times $\{t_{ij} | i = 1, \ldots, n_j, j = 1, \ldots, m\}$, and we label these ordered arrival times as $t_{(1)}, \ldots, t_{(n)}$. Note that here $n = \sum_{j=1}^{m} n_j$ if there are no ties in the observed event times. However, we do not assume that there are no ties.

2.1. *Estimation Based on Nelson–Aalen Estimator*

Our first estimator of τ_2 is based on finding slopes of the Nelson–Aalen estimator $\hat{\Lambda}(t)$ between ordered observed event times, $\frac{\hat{\Lambda}(t_{(i+1)}) - \hat{\Lambda}(t_{(i)})}{t_{(i+1)} - t_{(i)}}$. These slopes approximate $\lambda(t) = \frac{d\Lambda}{dt}$.

The Nelson–Aalen estimator is

$$\hat{\Lambda}(t) = \sum_{t_{(i)} \leq t} \frac{d_i}{Y_i}. \tag{2.1}$$

Here, $t_{(i)}$ is the ith ordered event time, d_i is the number of events at $t_{(i)}$, and Y_i is the number of events that have yet to happen (at risk set for recurrent events). See Lawless and Nadeau (1995) for more details on the use of the Nelson–Aalen estimator in the context of recurrent events.

Because $\Lambda(t)$ changes from concave to convex at τ_2, we would like to find the location of the minimum slope of $\hat{\Lambda}(t)$, which is a natural estimator of τ_2. However, it is very sensitive to the natural variability in the data. This is because the smallest slope would be chosen over single, adjacent event times. As a cure for this problem, we propose minimizing these slopes, taken over multiple adjacent intervals simultaneously. Then, the estimator can be found by averaging all time points inside the union of these adjacent intervals. If we wish to consider k intervals, then we can use

$$t_{(i^*)}^{k} = \arg\min_{t_{(i)}} \frac{\hat{\Lambda}(t_{(i+k)}) - \hat{\Lambda}(t_{(i)})}{t_{(i+k)} - t_{(i)}} \tag{2.2}$$

as the left endpoint of the union of adjacent intervals. Based on this, as our first estimator we propose

$$\hat{\tau}_2 = \frac{\sum_{v=0}^{k} t^k_{(i^*+v)}}{k+1}. \tag{2.3}$$

Note that in Equation (2.3) one can choose the value of k to minimize the variability in the estimate.

If it is believed that $\gamma_1(t)$ is increasing and $\gamma_2(t)$ is decreasing, then we define

$$t^k_{(i^*)} = \arg\max_{t_{(i)}} \frac{\hat{\Lambda}(t_{(i+k)}) - \hat{\Lambda}(t_{(i)})}{t_{(i+k)} - t_{(i)}}$$

with $\hat{\tau}_2$ the same as in Equation (2.3).
Consistency of this estimate is easy to establish, as shown in the following theorem.

Theorem 2.1. $\lim_{m\to\infty} \frac{\sum_{v=0}^{k} t^k_{(i^*+v)}}{k+1} = \tau_2$, *where* $t^k_{(i^*)} = \arg\min_{t_{(i)}} \frac{\hat{\Lambda}(t_{(i+k)})-\hat{\Lambda}(t_{(i)})}{t_{(i+k)}-t_{(i)}}$.

Proof We will proceed by showing that each $t^k_{(i^*+v)} \to \tau_2$ for any v and k. Without loss of generality, suppose that $\gamma_1(t)$ is decreasing and $\gamma_2(t)$ is increasing. Then,

$$\begin{aligned}\lim_{m\to\infty} \arg\min_{t_{(i)}} \frac{\hat{\Lambda}(t_{(i+k)}) - \hat{\Lambda}(t_{(i)})}{t_{(i+k)} - t_{(i)}} &= \arg\min_{t_{(i)}} \lim_{m\to\infty} \frac{\hat{\Lambda}(t_{(i+k)}) - \hat{\Lambda}(t_{(i)})}{t_{(i+k)} - t_{(i)}} \\ &= \arg\min_{t_{(i)}} \lim_{m\to\infty} \frac{\Lambda(t_{(i+k)}) - \Lambda(t_{(i)})}{t_{(i+k)} - t_{(i)}} \\ &= \arg\min_{t_{(i)}} \lambda(t_{(i)}) \qquad (2.4) \\ &\to \quad \tau_2. \qquad (2.5)\end{aligned}$$

Note that Equation (2.4) is true because we are assuming that there are no simultaneous events, so that $\lim_{m\to\infty} t_{(i+k)} - t_{(i)} = 0$. Equation (2.5) is true because the fact that $\gamma_1(t)$ is decreasing and $\gamma_2(t)$ is increasing guarantees that the location of the minimum will converge to the correct location. This completes the proof. □

2.2. *Estimation Based on Modified Version of Nelson–Aalen Estimator*

It is clear that even for datasets with many events, the increments in the estimator (2.1) are larger for larger values of t. This is because the risk set is smaller near the end of the time on study. Thus, events are not given equal weight. In fact, there can be large differences, and we expect this to affect our proposed estimator in Equation (2.3).

To correct this, we propose to effectively shrink the Nelson–Aalen estimator by inflating the denominator of the increments by a constant. We will require that the constant go to zero as the sample size goes to infinity. This correction should be chosen to be of the same order or larger as the risk set at the beginning of the time scale (total number of events) for smaller sample sizes. This would have the effect of more evenly weighting the events. For larger sample sizes, where the Nelson–Aalen estimator, $\hat{\Lambda}(t)$, has converged sufficiently close to the true cumulative intensity function, $\Lambda(t)$, the correction will be less necessary.

One way to do this is to choose

$$C_m = (C_{(1)} - t_0) a_m \hat{\Lambda}(C_{(1)} - t_0), \tag{2.6}$$

where $C_{(1)}$ is the smallest censoring time, t_0 is the common start time under study, and

$$a_m = \begin{cases} m & m \leq b \\ b\exp(-0.01(m-b)) & m > b. \end{cases}$$

Because $\hat{\Lambda}(t)$ is estimating the expected number of events per person, we multiply by the number of individuals, as well as the length of time, to obtain the desired result. Choosing a_m this way allows the meaningful correction of $\hat{\Lambda}(t)$ for smaller sample size while still forcing C_m to go to zero as $m \to \infty$. Simulations (see Section 3) show that $b = 200$ will correct the Nelson–Aalen estimator for smaller sample sizes, but also that there is not much to be gained for larger values of b in terms of bias and mean squared error. It should be noted that there surely are other ways to choose weights to facilitate this correction, however as explained, this method has intuitive appeal.

Thus, we define the corrected Nelson–Aalen estimator,

$$\hat{\Lambda}^{C_m}(t) = \sum_{t_{(i)} \leq t} \frac{d_i}{Y_i + C_m}. \tag{2.7}$$

Then, we can redefine $t^k_{(i^*)}$ in Equation (2.2) by

$$t^k_{(i^*)} = \arg\min_{t_{(i)}} \frac{\hat{\Lambda}^{C_m}(t_{(i+k)}) - \hat{\Lambda}^{C_m}(t_{(i)})}{t_{(i+k)} - t_{(i)}}. \tag{2.8}$$

Now, an alternative estimator of τ_2 can be obtained simply by replacing $t^k_{(i^*)}$ in Equation (2.3) with $t^k_{(i^*)}$ from Equation (2.8).

As long as $\hat{\Lambda}^{C_m}(t)$ converges to the true $\Lambda(t)$, then averaging the times over the interval giving the minimum (or maximum) slope still produces a consistent estimate of τ_2. In Theorem 2, we argue that the corrected estimator converges uniformly to the uncorrected Nelson–Aalen estimator. The properties of the Nelson–Aalen estimator are well known (see Andersen et al. 1993).

Theorem 2.2. *If* $\lim_{m\to\infty} C_m = 0$, *then* $\lim_{m\to\infty} \sup_{t\in[\tau_1,\tau_u]} |\hat{\Lambda}^{C_m}(t) - \hat{\Lambda}(t)| = 0.$

Proof: It is clear that

$$\begin{aligned} \sup|\hat{\Lambda}^{C_m}(t) - \hat{\Lambda}(t)| &= \sup|\sum_{t_{(i)}\leq t} \frac{d_i}{Y_i + C_m} - \sum_{t_{(i)}\leq t} \frac{d_i}{Y_i}| \\ &\leq \sup \sum_{t_{(i)}\leq t} |\frac{d_i}{Y_i + C_m} - \frac{d_i}{Y_i}| \\ &= \sup \sum_{t_{(i)}\leq t} \left(\frac{C_m}{Y_i + C_m} \frac{d_i}{Y_i} \right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{t_{(i)} \le \tau_u} \left(\frac{C_m}{Y_i + C_m} \frac{d_i}{Y_i} \right) \\
&\le \frac{C_m}{m + C_m} \hat{\Lambda}(\tau_u) \\
&\to 0.
\end{aligned}
$$

This completes the proof. □

In Theorem 2, τ_1 is as defined in Equation (1.2), and τ_u is a known upper bound on the unknown change-point, τ_2. Typically, a practitioner will have a preconceived notion about a value for τ_u. This is a standard restriction when estimating an unknown change-point.

Remark 2.1. Asymptotic distributions of our proposed estimators are very complicated and difficult to obtain. Throughout the paper, we have chosen to use bootstrap methods to construct confidence intervals for the change-point.

3. Simulations

In this section, through simulations, we describe how to implement our methodologies. We also compare our estimators. For our purposes, event times are generated between 0 and 17, with censoring times randomly chosen between 15 and 17. Without loss of generality, τ_1 is taken to be zero and τ_2 can be either 5 or 10.

The rates, $\gamma_1(t)$ and $\gamma_2(t)$ are taken to be linear, and we allow $\gamma_1(t)$ to be a decreasing function and $\gamma_2(t)$ to be an increasing function first, and then vice versa. The slope for $\gamma_1(t)$, denoted by λ_1, is always ± 0.05, while the slope for $\gamma_2(t)$, denoted by λ_2, is taken from the set $\mp\{0.0125, 0.025, 0.05, 0.10, 0.15\}$. The overall rate function, $\lambda(t)$, is always constructed to maintain continuity.

We generate 1,000 datasets for each parameter combination, and sample sizes of 20, 50, and 100 are used for each combination. For $m = 20$, we compute the estimate for $k = 1, \ldots, 150$. For $m = 50$, we use $k = 1, \ldots, 300$. Finally, for $m = 100$, we try $k = 1, \ldots, 450$. Though it is possible that individual datasets may not have enough events to support some of the larger values of k, typically the optimal value will occur before this is a problem. The other possibility is that the estimator will struggle to detect the change-point in a few cases, and the bias and MSE will decrease monotonically as k increases, due to averaging over more and more intervals.

When it comes to choosing the optimal value of k, we pick those 5–10 values that have the lowest bias, and then choose the minimum mean squared error (MSE) among these. This is because only using minimum MSE as the criterion can lead to marginally higher bias. When analyzing a real dataset, we will choose the value of k that has the lowest bootstrap estimate of variability (see the next section). In the tables that follow, we summarize our results.

If the change-point is early in the study, and the rate function is at first decreasing and then switches to increasing as in Table 1, we recommend the uncorrected estimator. However, if the change-point is expected to be later in the study (Table 2), then the corrected version performs better. In Table 3, we handle the case where the change-point is early and the rate switches from increasing $\gamma_1(t)$ to decreasing $\gamma_2(t)$. In this case, the corrected estimator performs better, though neither does extremely well. Finally, if the change-point is late and the rate starts out increasing and changes to decreasing (Table 4), there is not a

Table 1
Decreasing then increasing rate 1

$\tau_2 = 5, \lambda_1 = -0.05$		m	20	50	100
$\lambda_2 = 0.05$	Uncorrected	Bias ($\hat{\tau}_2$)	0.0033	0.0084	0.0983
		RMSE ($\hat{\tau}_2$)	0.4999	0.3028	0.2239
	Corrected	Bias ($\hat{\tau}_2$)	0.0044	0.0024	0.0014
		RMSE ($\hat{\tau}_2$)	0.9580	0.5445	0.3691
λ_2=0.025	Uncorrected	Bias ($\hat{\tau}_2$)	0.0046	0.0057	0.0387
		RMSE ($\hat{\tau}_2$)	0.5189	0.3064	0.2030
	Corrected	Bias ($\hat{\tau}_2$)	0.5528	0.2677	0.1696
		RMSE ($\hat{\tau}_2$)	2.0860	1.5484	1.3120
$\lambda_2 = 0.0125$	Uncorrected	Bias ($\hat{\tau}_2$)	0.0129	0.0066	0.0098
		RMSE ($\hat{\tau}_2$)	0.5459	0.3380	0.2126
	Corrected	Bias ($\hat{\tau}_2$)	1.5730	1.6448	2.9151
		RMSE ($\hat{\tau}_2$)	1.6369	1.9167	2.9251

clear winner. If the slope of the decreasing piece is very steep, the uncorrected procedure is better, but the corrected estimator should be used otherwise.

It is worth noting that the bias in some of the tables does not always decrease monotonically. This is because, in our choices of values for the parameters, we have been conservative. Using small values for the rates does not allow many events, despite having what seems to be reasonably large sample sizes.

4. Analysis of Example

A randomized double-blinded placebo-controlled study was performed by Dibley et al. (1996) to assess the effectiveness of Vitamin A supplements to reduce the incidence of

Table 2
Decreasing then increasing rate 2

$\tau_2 = 10, \lambda_1 = -0.05$		m	20	50	100
λ_2=0.05	Uncorrected	Bias ($\hat{\tau}_2$)	3.5112	3.6939	2.9630
		RMSE ($\hat{\tau}_2$)	3.5371	3.7093	4.2988
	Corrected	Bias ($\hat{\tau}_2$)	0.2065	0.0938	0.1712
		RMSE ($\hat{\tau}_2$)	1.6793	1.0863	0.7979
$\lambda_2 = 0.10$	Uncorrected	Bias ($\hat{\tau}_2$)	3.5327	2.7924	1.9241
		RMSE ($\hat{\tau}_2$)	4.5419	3.8997	2.7945
	Corrected	Bias ($\hat{\tau}_2$)	0.9971	0.6900	0.4595
		RMSE ($\hat{\tau}_2$)	1.9608	1.3103	0.9014
$\lambda_2 = 0.15$	Uncorrected	Bias ($\hat{\tau}_2$)	3.0893	2.2830	1.6253
		RMSE ($\hat{\tau}_2$)	3.2805	3.2377	2.1178
	Corrected	Bias ($\hat{\tau}_2$)	1.2042	0.8162	0.6162
		RMSE ($\hat{\tau}_2$)	1.9579	1.2981	0.9229

Table 3
Increasing then decreasing rate 1

$\tau_2 = 5, \lambda_1 = 0.05$		m	20	50	100
λ_2=−0.05	Uncorrected	Bias ($\hat{\tau}_2$)	0.1068	2.6859	3.6107
		RMSE ($\hat{\tau}_2$)	3.3145	2.7316	3.6196
	Corrected	Bias ($\hat{\tau}_2$)	0.0731	0.4465	0.4249
		RMSE ($\hat{\tau}_2$)	2.7903	1.4952	1.0707
λ_2=−0.025	Uncorrected	Bias ($\hat{\tau}_2$)	3.3433	4.1588	4.8984
		RMSE ($\hat{\tau}_2$)	3.4546	4.1699	4.9024
	Corrected	Bias ($\hat{\tau}_2$)	2.2541	2.0400	1.9227
		RMSE ($\hat{\tau}_2$)	3.3986	3.1046	2.7397
λ_2=−0.0125	Uncorrected	Bias ($\hat{\tau}_2$)	4.0506	4.7287	5.3594
		RMSE ($\hat{\tau}_2$)	4.0947	4.7374	5.3626
	Corrected	Bias ($\hat{\tau}_2$)	3.3733	3.9526	4.2851
		RMSE ($\hat{\tau}_2$)	3.5310	4.0621	4.4124

acute respiratory illness (ARI) in children. The authors also studied the impact on acute lower respiratory illness as well as diarrhea, but we limit our analysis to ARI.

An occurrence of ARI was defined to be two or more adjoining days for which a child was reported to have a cough. We take the start of each event to be our recurrent event. Though this means a child is not at risk for another event until the current bout of cough subsides, most episodes ended in a matter of days. In the context of a two-year study, these durations are taken to be negligible.

Thirty-four villages in Indonesia were surveyed, and 1,036 children aged six to 47 months were identified for the study, with more than 90% completing treatment. Some of these did not enter into the study until the second year, but the time scale used in the

Table 4
Increasing then decreasing rate 2

$\tau_2 = 10, \lambda_1 = 0.05$		m	20	50	100
λ_2=−0.05	Uncorrected	Bias ($\hat{\tau}_2$)	0.3378	0.7500	1.1175
		RMSE ($\hat{\tau}_2$)	0.5325	0.7850	1.1262
	Corrected	Bias ($\hat{\tau}_2$)	0.0005	0.2645	0.6786
		RMSE ($\hat{\tau}_2$)	0.8270	0.6975	0.8500
λ_2=−0.10	Uncorrected	Bias ($\hat{\tau}_2$)	0.0088	0.2460	0.7008
		RMSE ($\hat{\tau}_2$)	0.4197	0.3429	0.7164
	Corrected	Bias ($\hat{\tau}_2$)	0.2542	0.0993	0.0247
		RMSE ($\hat{\tau}_2$)	1.4313	1.0607	0.8188
λ_2=−0.15	Uncorrected	Bias ($\hat{\tau}_2$)	0.0087	0.0054	0.0962
		RMSE ($\hat{\tau}_2$)	0.4431	0.2668	0.2127
	Corrected	Bias ($\hat{\tau}_2$)	0.7691	0.4991	0.4000
		RMSE ($\hat{\tau}_2$)	1.5846	1.0475	0.8933

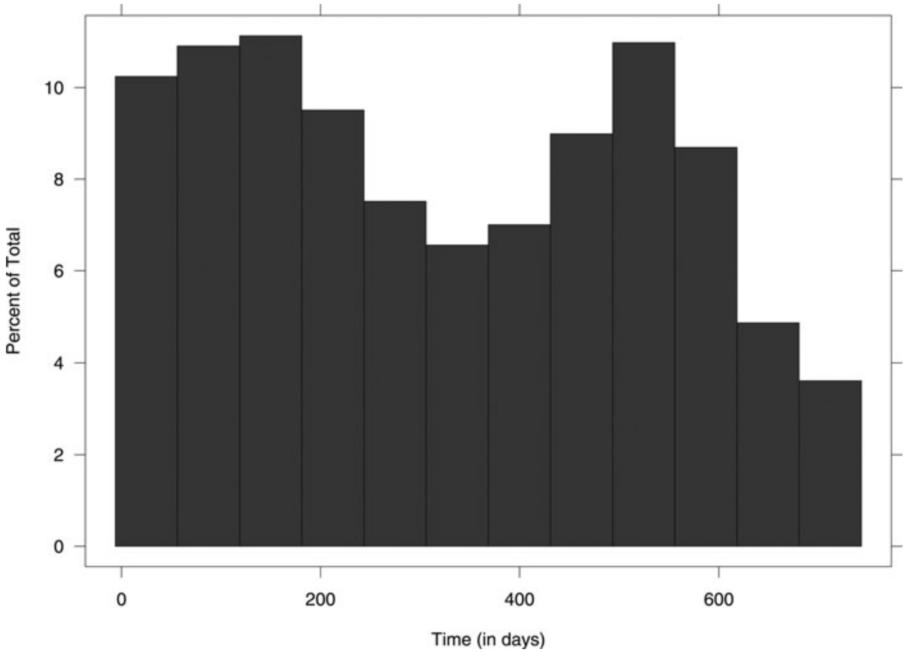

Figure 1. Time to incidence of acute respiratory illness.

analysis accounts for this. Blocking was used to ensure balance of groups by village and distance from health services.

Children were randomly assigned to receive a vitamin A supplement or placebo every 4 months for 2 years. Though this design may suggest the possibility of multiple unknown changepoints, we use a model with only one, occurring at some time after the initial treatment. The subsequent treatments could be thought to only maintain the level of vitamin A in patients.

A total of 10,735 episodes of ARI were observed, with 88% of children experiencing two or more. Interestingly, the incidence of ARI was 8% higher in the treatment group. The investigators theorize that the immune systems of children with preexisting sufficient levels of vitamin A are suppressed when levels of vitamin A are too high, leading to more frequent infections. Thus, we focus on the children in the poor nutrition group, as the investigators believe that other subjects would not be appropriate. Also, we require individuals to be under observation for at least 1.5 of the 2 years in the original study design, so that each included child contributes to the estimation of the unknown change-point. These restrictions yield a sample of size 102.

The main question we answer is, when does the effect of the Vitamin A supplement wear out? That is, when do we expect the supplement to stop suppressing the incidence of ARI, if not permanently? Is there a time at which the supplement starts to lose its effectiveness?

From Figure 1, the histogram of the event times, it is clear that the data follow a decreasing and then increasing rate function.

We would like to detect the time when the rate starts to increase. We hope to detect the location of this turning point, as it may be the time at which the Vitamin A treatment ceases to be effective. The histogram suggests that this occurs around 1 year, or 365 days.

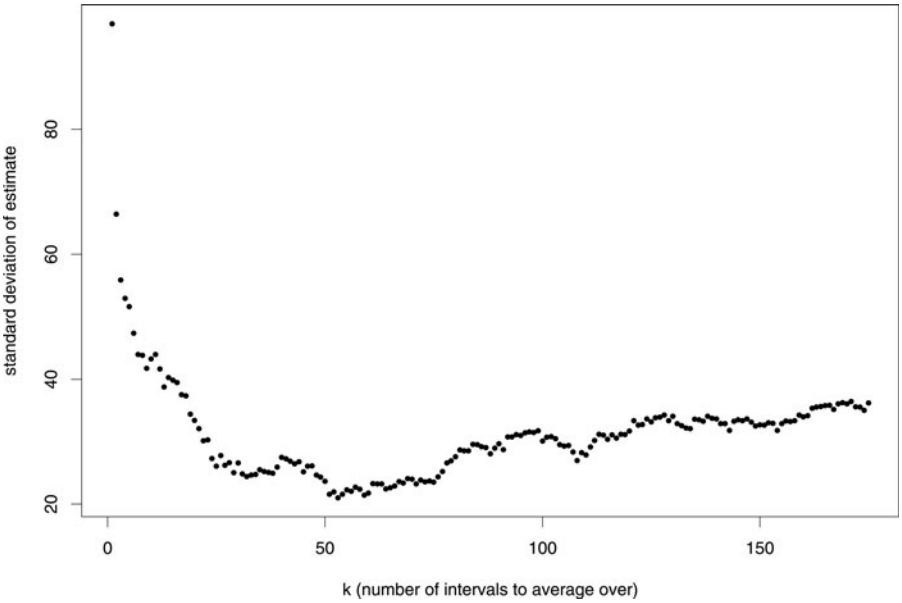

Figure 2. Variability of estimate as a function of *k*.

In the analysis, we use 120 days as a lower bound for the unknown τ_2 because patients receive a dose of Vitamin A every 4 months. Also, 540 days is chosen as an upper bound, as that is when we start to see patients leaving the study.

In order to determine the optimal value of k, we generated 250 bootstrap resamples, and computed the standard deviation of $\hat{\tau}_2$ for each value of k. The resampling was done by randomly choosing whole subjects rather than individual events, in order to maintain independence between units. We only check up to 175 for k, because many of the resamples will start to run out of events after that point. Figure 2 shows how averaging over multiple intervals can drastically reduce the variability of the estimate.

The minimum of 21.0225 occurs at $k = 53$. Using $k = 53$, we get an estimate for τ_2 of 361.8333, which agrees with the histogram. It is interesting to note that we get the same value regardless of if we choose the lower bound to be 0 or 120. This is because our method is not directly dependent on τ_1.

As we mentioned in Remark 2, since asymptotic distributions of our proposed estimators are very complicated, we use the hybrid bootstrap method (Shao and Tu 1996) to obtain a 95% confidence interval for τ_2. Intervals based on the normal approximation as well as the bootstrap percentile were considered, but there were differences, and the hybrid method has better coverage probability (Shao and Tu 1996). Using their notation, $H_{\text{boot}}^{-1}(1-0.025) = 291.7825$ and $H_{\text{boot}}^{-1}(1-0.975) = -619.5622$ for 2,500 resamples. This yields a hybrid bootstrap 95% confidence interval for τ_2 of (332.9, 423.2).

For the sake of argument, the uncorrected estimate was also computed. The optimal value of k, 40, yielded a point estimate of $\hat{\tau}_2 = 311.4139$ and a standard error of 54.0122. The histogram of event times suggests that this estimate is biased. Also, note that the standard error is much higher for this uncorrected estimator. This is consistent with the results of the simulations in the previous section. It appears that the change-point is not early in the study, so we should choose the corrected version.

5. Concluding Remarks

Two estimators of an unknown change-point (τ_2) are proposed in the context of recurrent events. Though the given formulation includes a known change-point τ_1 (e.g., for time of intervention or treatment), such a feature is not central to the use of the estimators. However, we do require that a lower (τ_1) and upper bound (τ_u) are known for τ_2.

A major advantage of the proposed methods is that no parametric assumptions are made about the intensity rate, $\lambda(t)$, and by implication, the cumulative intensity function, $\Lambda(t)$. The unknown change-point, τ_2, separates the post-treatment intensity rate into increasing and decreasing components. This is the only restriction we put on $\lambda(t)$.

Both estimators are based on computing slopes of the Nelson–Aalen estimator of the cumulative intensity function over intervals between observed event times. In order to minimize the variability of either estimate, we compute slopes over multiple adjacent intervals. Because the increments in the Nelson–Aalen estimator are always larger for later events, the second proposed estimator shrinks the increments to more equally weight the events. Both estimators are shown to be consistent, but the second estimator performs better in terms of bias and mean squared error for most cases tested. A natural extension of the model (1.2) is to include covariates in the model, and this is currently under investigation.

References

Aschar, J. A., Loibel, S., Andrade, M. (2007). Interfailure data with constant hazard function in the presence of change-points. *REVSTAT* 5:209–226.

Andersen, P., Borgan, O., Gill, R., Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Cook, R. J., Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. New York: Springer.

Dibley, M. J., Sadjimin, T., Kjolhede, C. L., Moulton, L. H. (1996). Vitamin A supplementation fails to reduce incidence of acute respiratory illness and diarrhea in preschool-age Indonesian children. *Journal of Nutrition* 126:434–442.

Frobish, D., Ebrahimi, N. (2007). Parametric estimation of change-points for actual event data in recurrent events models. *Computational Statistics and Data Analysis* 53:671–682.

Gonzalez, J. R., Pena, E. A., Delicado, P. (2010). Confidence intervals for median survival with recurrent event data. *Computational Statistics and Data Analysis* 54:78–89.

Karasoy, D. S., Kadilar, C. (2007). A new Bayes estimate of change-point in the hazard function. *Computational Statistics and Data Analysis* 51:2993–3001.

Lawless, J., Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* 37:158–168.

Matthews, D., Farewell, V. (1982). On testing for a constant hazard against a change-point alternative. *Biometrics* 38:463–468.

Pena, E., Slate, E., Gonzalez, J. (2007). Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference* 137:1727–1747.

Pena, E., Stocker, R. (2007). A general class of parametric models for recurrent event data. *Technometrics* 49:210–221.

Shao, J., Tu, D. (1996). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Wu, C. Q., Zhao, L. C., Wu, Y. H. (2003). Estimation in change-point hazard function models. *Statistics and Probability Letters* 63:41–48.

Zhang, W., Qian, L., Li, Y. (2014). Semiparametric sequential testing for multiple change points in piecewise constant hazard functions with long-term survivors. *Communications in Statistics* 43:1685–1699.