Published in final edited form as:

Stat Med. 2010 March 30; 29(7-8): 896–905. doi:10.1002/sim.3808.

Threshold Regression for Survival Data with Time-varying Covariates

Mei-Ling Ting Lee^{1,*}, G. A. Whitmore², and Bernard Rosner³

¹ Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD, U.S.A

² McGill University, Montreal, Quebec, Canada

³ Brigham and Women's Hospital and Harvard Medical School, Boston, MA, U.S.A

SUMMARY

Time-to-event data with time-varying covariates pose an interesting challenge for statistical modeling and inference, especially where the data require a regression structure but are not consistent with the proportional hazard assumption. Threshold regression (TR) is a relatively new methodology based on the concept that degradation or deterioration of a subject's health follows a stochastic process and failure occurs when the process first reaches a failure state or threshold (a first-hitting-time). Survival data with time-varying covariates consist of sequential observations on the level of degradation and/or on covariates of the subject, prior to the occurrence of the failure event. Encounters with this type of data structure abound in practical settings for survival analysis and there is a pressing need for simple regression methods to handle the longitudinal aspect of the data. Using a Markov property to decompose a longitudinal record into a series of single records is one strategy for dealing with this type of data. This study looks at the theoretical conditions for which this Markov approach is valid. The approach is called *threshold regression with Markov decomposition* or *Markov TR* for short. A number of important special cases, such as data with unevenly spaced time points and competing risks as stopping modes, are discussed. We show that a proportional hazards regression model with time-varying covariates is consistent with the Markov TR model. The Markov TR procedure is illustrated by a case application to a study of lung cancer risk. The procedure is also shown to be consistent with the use of an alternative time scale. Finally, we present the connection of the procedure to the concept of a collapsible survival model.

Keywords

Competing risks; first hitting time; latent process; longitudinal data; Markov property; stopping time; unevenly spaced time points; Wiener diffusion process

1. Introduction

Longitudinal data in survival analysis refer to sequential observations on the level of degradation and/or time-varying covariates of the subject, prior to the occurrence of the failure event. Encounters with this type of data structure abound in practical settings and there is a pressing need for regression methods to handle the longitudinal aspect, especially where the survival functions do not possess the proportional hazards property. Lee and Whitmore [1] review a relatively new regression methodology for survival data referred to

*Correspondence to: MLTLEE@UMD.EDU.

as *threshold regression* or *TR* for short. The methodology is based on the concept that degradation or deterioration of a subject's health follows a stochastic process and failure occurs when the process first reaches a failure state or threshold (a first hitting time). They mention the case of longitudinal data with time-varying covariates but do not investigate it in depth.

Decomposing or splitting a longitudinal record into a series of single records with only initial and closing measurements in each record is a strategy for dealing with longitudinal data. The idea has been considered by other researchers but not studied fully. For example, the proposal by Efron [3] for parametric regression modeling for hazard rates and survival curves is in this spirit and could be extended to include time-varying covariates. D'Agostino *et al* [2] compared the effect of pooling logistic regressions performed at each time point with proportional hazards (PH) regression having time-dependent covariates. In this article, we look at the formal conditions that must hold for this decomposition of a longitudinal record to be a valid procedure in the context of TR. The conditions are examined in terms of both theory and practical application. The key property that must hold is a Markov property, as we will show later. We therefore refer to this procedure as *threshold regression with Markov decomposition* or *Markov TR* for short. Statistical software packages have data routines that can perform this decomposition automatically (for example, *stsplit* in Stata).

Section 2 presents a brief overview of the TR model. Section 3 then introduces the formalities of a longitudinal process in which both the underlying health process and covariate processes are observable. Section 4 presents the theoretical justification for Markov decomposition. Section 5 presents the partial likelihood approach to estimation and inference for the Markov TR approach. Section 6 presents theory for the important practical setting where the underlying health process is unobservable but the covariate processes remain observable. Section 7 enumerates theoretical and practical results for some further special cases, such as competing risks as stopping modes and data with unevenly spaced time points. The connection between Markov TR and proportional hazards regression is examined in Section 8. A case illustration involving lung cancer risk is presented in Section 9. In Section 10, we show that the Markov TR procedure can be used effectively in conjunction with an alternative time scale and we link the procedure to the concept of a collapsible survival model. Finally, Section 11 presents a closing review and discussion.

2. First-hitting-time Based Threshold Regression Models

Threshold regression (TR) refers to a statistical model in which the event time S is the *first-hitting-time* (FHT) of an absorbing boundary or *threshold* B by a stochastic process $\{Y(t), t \geq 0\}$. In other words,

$$S = \inf \{t: Y(t) \in B\}. \quad (1)$$

A regression structure is introduced by having parameters of the model depend on covariates through appropriate link functions. See Aalen *et al* [4] for a review of first hitting time models.

Various terminology is used for the elements in the TR formulation. The event of interest may refer to death, failure, relapse or another medical endpoint. The time to the event itself may be called a survival time, lifetime, failure time, event time or other similar term. The data analysis may be referred to as survival data analysis, event history analysis or something similar. The practical context may be health and medicine (as in this article), engineering, the natural sciences, economics, or the social sciences. The alternative terms

are not exact synonyms but are roughly equivalent. The concepts and methods in this article are applicable to all of these contexts.

To illustrate the TR ideas, consider a study of chronic liver disease. The disease tends to progress in a subject through various discrete stages from disease-free (stage 0), inflammation (stage 1), fibrosis (stage 2), cirrhosis (stage 3), and complete liver failure (stage 4) which leads inescapably to either death or, possibly, a liver transplant. Entry into stage 4 may define the endpoint of interest. The disease stage of a subject is observed by a clinician at regular or irregular clinical visits. The stage of disease observed at each visit represents a longitudinal reading on the health process $\{Y(t)\}$ in the TR model. Stage 4 is an absorbing state or boundary B . The time of entry to stage 4 is the first hitting time S of the model. The covariates refer to characteristics of each subject or the subject's environment (such as age, gender, physiological factors, diet, medical treatments, and environmental exposures) as these characteristics vary over the course of the study and influence the health of the subject. This medical setup might be described by a semi-Markov process with five states, the last of which is an absorbing state (stage 4). As a second illustration, consider our case application of the TR model that is discussed in Section 9. In this application, the health process of each subject refers to lung cancer and is considered a latent or unobservable process. Its form is assumed to be a Wiener diffusion process $\{Y(t)\}$ where t denotes time measured from the start of an observation interval. The endpoint is a diagnosis of primary lung cancer and occurs when the health process first decreases to the zero level, which is taken as the threshold for the event (boundary B). The parameters of the Wiener process are connected to covariates (such as age and smoking history) by regression link functions.

Lee, Whitmore, and Rosner [5] demonstrate a univariate TR model at a single time point and present a comparison with Cox proportional hazards regression. Lee, DeGruttola and Schoenfeld [6] use a bivariate Wiener model for survival data to represent a latent health process and a correlated marker process. These authors mention an interesting approach to handling longitudinal data that they anticipated would be technically satisfactory and practical to implement. Their suggested approach, however, is not developed in their article. Lee and Whitmore [1] elaborate somewhat on this approach but leave a full exploration of the approach as an open research question. They refer to this approach as the *uncoupling* procedure. This paper takes up the elaboration called for in the earlier article.

3. Theory for a Fully Observable Longitudinal Process

Applications of TR are distinguished by whether observations are available on the actual health process $\{Y(t)\}$ or whether this process remains unobservable or latent. Observable readings on health status tend to arise where health is described by a clinical construct such as clinically measurable or recordable symptoms, recorded medical outcomes, or complex combinations of these observables. The five stages of chronic liver disease is an illustration of an observable health process – observable because health status is defined by clinically defined disease stages that are agreed upon or adjudicated by specialists. Thus, a health process is an observed process according to whether an investigator chooses to define the health index of interest in terms of measurable quantities and/or observable entities.

In this section, we explore situations with longitudinal observations on *both* the health status and covariates of a subject. We denote the sequence of fixed time points at which observations are available for a subject by $t_0 \leq t_1 \leq \dots$. These time points need not be equally spaced. The longitudinal observation sequence ends when either the subject fails or the sequence is censored. We let C denote the censoring time and assume that, if censoring occurs, it occurs at one of the time points t_j . The failure and censoring times define two sequences of indicator variables: (1) a failure code sequence $\{f_j, j = 0, 1, \dots\}$, with $f_j = 0$ if S

$> t_j$ and $f_j = 1$ if $S \leq t_j$ and (2) a censoring code sequence $\{c_j, j = 0, 1, \dots\}$, with $c_j = 0$ if $C > t_j$ and $c_j = 1$ if $C = t_j$. With these definitions, the two codes have three possible configurations: (1) $f_j = 0, c_j = 0$, (2) $f_j = 1, c_j = 0$, and (3) $f_j = 0, c_j = 1$. The first configuration implies that the subject is surviving and uncensored at time t_j . The second implies that the subject is observed to fail before time t_j . The third implies that the subject survives beyond time t_j but the sequence is censored at that time. Thus, the last two configurations stop the longitudinal sequence. We assume that the subject is neither censored nor failed at the outset of the sequence; thus, $f_0 = c_0 = 0$.

In this formulation, we assume the process $Y(t)$ is observable at each time point. We denote this sequence of observations on the process level by $\{y_j, j = 0, 1, \dots\}$. We allow for observed covariates. These may be fixed or time varying. We denote the sequence of covariate vectors by $\{z_j, j = 0, 1, \dots\}$. If a sequence is censored at time t_m , so $c_m = 1$, we assume its process level y_m and covariate vector z_m are still observed. If the subject has failed at time t_m , so $f_m = 1$, then the process has entered the boundary set B at time S where $t_{m-1} < S \leq t_m$ but the value of y_m may or may not be defined. Likewise, some components of the covariate vector z_j may or may not be defined, depending on whether they are external or internal covariates (see the definitions in Kalbfleisch and Prentice [7], pages 122–127).

4. Decomposing a Longitudinal Record into Markov Transitions

As just noted, the longitudinal sequence of observations is stopped by occurrence of censoring or failure, whichever comes first. We let $j = m$ define the last observation in the random sequence. Thus, $j = m$ at the first observation for which either $f_m = 1$ or $c_m = 1$. We next define events $A_j = (y_j, z_j, f_j, c_j), j = 0, 1, \dots, m$. Then the observation sequence is a realization of the stopped stochastic process $\{A_j, j = 0, 1, \dots, m\}$. The probability of the observation sequence can be expanded in a standard way as a product of conditional probabilities as follows.

$$P(A_m, A_{m-1}, \dots, A_0) = P(A_0) \prod_{j=1}^m P(A_j | A_{j-1}, \dots, A_0) \quad (2)$$

We now make the key assumption that the stopped stochastic process $\{A_j, j = 0, 1, \dots, m\}$ is a Markov process. The Markov assumption allows us to simplify (2) as follows.

$$P(A_m, A_{m-1}, \dots, A_1) = P(A_0) \prod_{j=1}^m P(A_j | A_{j-1}) \quad (3)$$

In other words, the probability of observing A_j depends only on its preceding state A_{j-1} and not on the earlier history of the observation process. Statement (3) provides the theoretical justification for the *Markov decomposition procedure*. Observe that (3) amounts to splitting the longitudinal record into a series of single-step records. Record A_{j-1} sets initial conditions at the start of the j th time interval $(t_{j-1}, t_j]$ while record A_j describes the outcome at the end of the interval. Probability $P(A_j | A_{j-1})$ is the transition probability for the Markov observation process over the j th time interval. By breaking the observation sequence into a set of individual one-step Markov transitions, the longitudinal feature is effectively removed and longitudinal data become as easy to handle for statistical inference as cross-sectional data.

The Markov assumption is widely encountered in statistics and, fortunately, is often realistic as well. The property has played a major role in survival data analysis (see, for example, Hougaard [8] for an extensive collection of ideas and methods). Of course, assuming that the Markov property holds is easier than actually demonstrating it. Thus, our proposed approach requires validation of the Markov property. In many practical settings, it is found that the current condition of a subject encapsulates all useful information from the subject's history for inferring the forward trajectory of the subject. Thus, to the extent that the process level y_{j-1} and the covariate vector z_{j-1} fully describe the condition of the subject at time t_{j-1} then the Markov assumption will hold and will allow valid probability statements to be made about the subject's future condition and possible failure at or before the next time point t_j . Where the Markov assumption does not quite hold, it may be possible to adapt or modify the observation process so it is valid. Expanding the state space of the observation process, for example, is one such modification.

The explicit form of the probability elements on the righthand side of (3) can be written as follows.

$$P(A_0)=P(y_0, z_0, f_0=0, c_0=0) \quad (4)$$

$$P(A_j|A_{j-1})=P(y_j, z_j, f_j, c_j|y_{j-1}, z_{j-1}, f_{j-1}=0, c_{j-1}=0) \text{ for } j=1, \dots, m \quad (5)$$

The probability $P(A_0)$ in (4) states in symbols that the observation sequence starts with the process level y_0 , the initial covariate vector z_0 , and the subject alive and uncensored. If the initial conditions are taken as determined (as opposed to being random) then $P(A_0) = 1$. The conditional probability $P(A_j|A_{j-1})$ in (5) is present on the righthand side of (3) only if the subject survives and the sequence is uncensored at time t_{j-1} . Hence, the notation shows explicitly that $f_{j-1} = c_{j-1} = 0$. The conditional probabilities continue to unfold on the righthand side of (3) until the sequence is stopped by failure or censoring. For the final conditional probability $P(A_m|A_{m-1})$, one or both of y_m and z_m may be undefined if $f_m = 1$.

5. Partial Likelihood

The joint probability in (3) represents the sample likelihood contribution of the observation sequence for any single subject. The typical conditional probability $P(A_j|A_{j-1})$ in (3) can be factored into the following product of two probabilities.

$$P(y_j, f_j|z_j, c_j, y_{j-1}, z_{j-1}, f_{j-1}=0, c_{j-1}=0)P(z_j, c_j|y_{j-1}, z_{j-1}, f_{j-1}=0, c_{j-1}=0) \quad (6)$$

An alternative factorization is discussed later as a special case.

Partial likelihood provides a basis for estimating parameters of the threshold regression model. We let θ denote the parameter vector of the threshold regression model. Thus far, we have suppressed θ in the notation. Now we partition θ into (θ_1, θ_2) . Parameters θ_1 govern the process $\{Y(t)\}$ and boundary set B , and parameters θ_2 govern the joint process $\{(z_j, c_j), j = 1, \dots\}$ for the covariate vector and censoring mechanism. We assume that θ_1 is of primary interest in the regression analysis. In many applications, the joint process for the covariate vector and censoring mechanism does not depend on parameters θ_1 . In this case, the contribution of the observation sequence to the sample partial likelihood function for θ_1

involves only the first term of (6). The contribution from a single observation sequence to the sample partial likelihood function then has the following form (after setting $P(A_0) = 1$):

$$L(\theta_1) = \prod_{j=1}^m P(y_j, f_j | z_j, c_j, y_{j-1}, z_{j-1}, f_{j-1}=0, c_{j-1}=0) \quad (7)$$

The sample partial likelihood function contains only conditional probability terms involving the process levels y_j and failure events $f_j, j = 1, \dots, m$. The likelihood contribution in (7) provides the basis for statistical inference about parameter vector θ_1 .

If the censoring mechanism is independent of the process $\{Y(t)\}$ then the c_j terms can be dropped from the entries A_j in the observation process $\{A_j, j = 0, \dots, m\}$ and the entries take the form $A_j = (y_j, z_j, f_j)$. The conditional probabilities on the righthand side of (7) now become:

$$P(y_j, f_j | z_j, y_{j-1}, z_{j-1}, f_{j-1}=0).$$

6. Theory for a Longitudinal Process with Unobservable Health Status

In many applications, the health process $\{Y(t)\}$ is taken as latent or unobservable. This is the case when an investigator views the development of the health condition of a subject as too complex or subtle to be easily measured or recorded. The lung cancer application in section 9 is such a case.

Where a study employs a latent health process, it provides no y_j readings. The entries A_j of the longitudinal observation process $\{A_j, j = 0, \dots, m\}$ take the form $A_j = (z_j, f_j, c_j)$. If we invoke the Markov assumption the y_j notation is dropped from righthand terms in (7), so each term has the form:

$$P(f_j | z_j, c_j, z_{j-1}, f_{j-1}=0, c_{j-1}=0)$$

Furthermore, if we assume independent censoring, the elements of the observation process become $A_j = (z_j, f_j)$ and each term in (7) has the following simple form:

$$P(f_j | z_j, z_{j-1}, f_{j-1}=0) \quad (8)$$

This version is one of the most commonly encountered in applications. The term will have one of the following two forms:

1. If $j \leq m$ and the subject is surviving at time t_j then:

$$P(f_j=0 | z_j, z_{j-1}, f_{j-1}=0) = P(S > t_j | z_j, z_{j-1}, S > t_{j-1}) \quad (9)$$

2. If $j = m$ and the subject has failed at time t_m then:

$$P(f_m=1 | z_m, z_{m-1}, f_{m-1}=0) = P(S \leq t_m | z_m, z_{m-1}, S > t_{m-1}) \quad (10)$$

In the first form (9), the conditioning involves only the initial and final covariate vectors, i.e., z_{j-1} and z_j , and survival to time t_{j-1} . In the form (10), the conditioning is the same except that the final covariate vector z_m may be undefined (because the subject has failed).

The latency of the health process and the consequent absence of readings y_j in this variant of Markov TR raises the issue of what information is lost by this absence. The loss may be less than anticipated because the covariate vector z_j can convey considerable information about the latent health level y_j . Indeed, the aim of TR modeling is to include covariates that are as informative as possible about the underlying health process. A useful surrogate or proxy for the latent process can be constructed from informative covariates as follows. The surrogate, denoted by h_j at time t_j , may be constructed as an estimate from a regression function of form $h_j = h(z_j)$. This idea was implemented by Whitmore, Crowder, and Lawless [9]. They refer to h_j as a *composite marker* for the latent health process. In the application in Section 9, we estimate the initial latent health level y_{j-1} for the j th observation interval by a log-linear regression function $\ln(y_{j-1}) = z_{j-1}\beta$, where β is a vector of regression coefficients estimated from the data. The estimate of y_{j-1} provided by this regression analysis serves as a surrogate health level h_{j-1} for time t_{j-1} . The more informative is z_{j-1} , the more precisely h_{j-1} imitates y_{j-1} .

7. Some Special Cases

A number of important special cases of the general Markov TR formulation arise in practical applications. We now itemize a few of these cases.

1. Fixed baseline covariates only

A special case arises if the data set has only fixed baseline covariates (i.e., no time-varying covariates). In this case, $z_j = z_0$ for all j . In addition, if no process observations y_j are available then the data set loses its longitudinal feature and the data can be analyzed using censored survival regression techniques, with baseline covariate vector z_0 . We mention this case because the Markov feature holds trivially; the conditional probability in (8) reduces to $P(A_j|A_{j-1}) = P(f_j|z_0, f_{j-1} = 0)$.

2. An alternative factorization

The factorization of $P(A_j|A_{j-1})$ in (6) is not unique. The following product is an alternative factorization that may be of practical value.

$$P(z_j, c_j|y_j, f_j, y_{j-1}, z_{j-1}, f_{j-1}=0, c_{j-1}=0)P(y_j, f_j|y_{j-1}, z_{j-1}, f_{j-1}=0, c_{j-1}=0) \quad (11)$$

As a general rule, the second term of (11) only depends on parameter vector θ_1 , but it may be less informative than the corresponding term in the original factorization (6) because the second probability in (11) is not conditioned on z_j and c_j . This factorization may be useful selectively. For example, for periods in which a subject fails, the value of z_j may be undefined and, hence, the previous factorization is meaningless. For the common case where the censoring mechanism is independent and the process $\{Y(t)\}$ is unobservable, then the rightmost probability in (11) reduces to $P(f_j|z_{j-1}, f_{j-1} = 0)$, which is a simple form.

3. Exact failure times

For some applications, it is reasonable to assume that the failure time of a failing item is exactly $S = t_m$ rather than being interval censored with $S \in (t_{m-1}, t_m]$. Then a probability

such as $P(S \leq t_m | z_m, z_{m-1}, S > t_{m-1})$ is replaced by the conditional probability density $g(\cdot)$ where

$$g(t_m | z_m, z_{m-1}, S > t_{m-1}) = \frac{dP(S \leq t_m | z_m, z_{m-1}, S > t_{m-1})}{dt_m}. \quad (12)$$

4. Varying time grids

In the Markov TR approach the time grid may have unevenly spaced points. The uneven spacing may derive from missed visits, record failures, administrative irregularities, and so on. The time grid can also differ from one subject to another without causing analytical difficulties for the method. Letting $t_{ij}, j = 0, 1, 2, \dots$ denote the grid for subject i , we denote the partition of the time line that includes the time points of all subjects by the set $\mathcal{T} = \{t_g, g = 0, 1, 2, \dots\}$. It is now clear that the longitudinal readings for any single subject may not be available for all time points in the common grid $\mathcal{T}$. The Markov assumption still holds without longitudinal readings. Some studies may require an analysis to span subjects across this common time grid. For example, a childhood obesity study may be concerned with shared social exposures through time as a covarying process. Computational procedures for Markov decomposition must be modified to accommodate this practical variant but no fundamental difficulties arise.

5. Using an alternative time scale

The decomposition procedure can be used in conjunction with an alternative operational time scale, such as cumulative exposure to a toxin. The resulting formulation then has links to the concept of a collapsible survival model. We explore this link in Section 10. If we let $r(t)$ denote the transformation of calendar time t to alternative time r , with $r(0) = 0$, and we let $\{Y^*(r)\}$ be the underlying health process defined in terms of alternative time r , the resulting subordinated health process $\{Y(t)\} = \{Y^*[r(t)]\}$ is the process $\{Y(t)\}$ defined earlier with respect to calendar time. The alternative time transformation may be a function of unknown parameters as well as covariates in vector $\mathbf{z}$. The alternative time transformation enters into the calculation of probability expressions that appear in the partial likelihood contribution (7) but otherwise does not alter the method of application of the Markov TR approach. The parameters of the alternative time transformation are included in the vector θ_1 along with other parameters of the underlying process $\{Y^*(r)\}$ and boundary set B .

6. Competing risks as stopping modes

The Markov decomposition procedure can be extended to competing risks if an M -variate indicator vector $\mathbf{m}_j$ is used to define the stopping event. The j th record then has the form $A_j = (y_j, z_j, \mathbf{m}_j)$. The event $\mathbf{m}_j$ equals $\mathbf{0} = (0, \dots, 0)$ if the process does not stop at time t_j and takes one of the unit values $\mathbf{1}_1 = (1, 0, \dots, 0)$, $\mathbf{1}_2 = (0, 1, \dots, 0)$, $\dots$, $\mathbf{1}_M = (0, 0, \dots, 1)$ if the process stops at time t_j . If the c th component of $\mathbf{m}_j$ is 1, the c th mode has stopped the process. Censoring can be included as a stopping mode, say, the M th mode. In the Markov TR model described here, $\mathbf{m}_j = (f_j, c_j)$. The two stopping modes are failure and censoring. This TR formulation allows the joint modeling and analysis of longitudinal data for competing risks, including the censoring mechanism.

8. Piece-wise Constant Proportional Hazards Model

We now demonstrate that the piece-wise constant proportional hazards (PH) regression model is consistent with the Markov TR model. The result shows that Markov TR is not an exotic idea but one that is in harmony with accepted methods for analyzing longitudinal data

and, in particular, time-varying covariate data. We denote the hazard function of the PH model by $h[t|z(t)] = h(t|0) \exp[z(t)\beta]$ where $z(t)$ is a time-varying covariate vector and $h(t|0)$, a baseline hazard function. The PH model does not require that its hazard functions arise from an underlying first hitting time. Yet, almost any family of hazard functions of the PH form can be constructed from selected combinations of a health process $\{Y^*(r)\}$ defined in terms of alternative time, a boundary set B , and an alternative time scale $r(t)$. In these cases, proportional hazards regression is a special case of threshold regression in which the form of the underlying process, boundary and time scale remain unspecified.

In our demonstration, the simple form of Markov decomposition is set out in (8) based on no process readings and an independent censoring mechanism. We first observe the following connection between the survival function and the hazard function for the PH model with piece-wise constant covariates and a piecewise baseline hazard function.

$$P(S > t_j | z_j, z_{j-1}, \dots, z_0) = \exp \left[- \sum_{g=1}^j h_{0g} \Delta t_g \exp(z_{g-1}\beta) \right] \quad (13)$$

Here the baseline hazard function is defined by $h(t|0) = h_{0g}$ for $t_{g-1} \leq t < t_g$ and $\Delta t_g = t_g - t_{g-1}$, for $g = 1, \dots, j$. For expository convenience, we have arbitrarily assumed that the covariate vector z_{g-1} remains constant over the (forward) time interval $[t_{g-1}, t_g)$. With the preceding setup, the conditional survival function becomes

$$\begin{aligned} P(S > t_j | z_j, z_{j-1}, \dots, z_0, S > t_{j-1}) &= P(S > t_j | z_{j-1}, S > t_{j-1},) \\ &= \exp[-h_{0j} \Delta t_j \exp(z_{j-1}\beta)]. \end{aligned} \quad (14)$$

The preceding formula implies that (9) has the following form in this case:

$$P(S > t_j | z_j, z_{j-1}, S > t_{j-1}) = \exp[-h_{0j} \Delta t_j \exp(z_{j-1}\beta)], \quad \text{for } j \leq m, \quad (15)$$

where z_j can be dropped because the piece-wise constant formulation chosen in (13) does not depend on z_j . Likewise, assuming interval censoring of failure, formula (10) has the form

$$P(S \leq t_m | z_m, z_{m-1}, S > t_{m-1}) = 1 - \exp[-h_{0m} \Delta t_m \exp(z_{m-1}\beta)]. \quad (16)$$

Finally, assuming that a failure occurs exactly at t_m when it occurs, then (12) has the following form in this case:

$$g(t_m | z_m, z_{m-1}, S > t_{m-1}) = h_{0m} \exp(z_{m-1}\beta) \exp[-h_{0m} \Delta t_m \exp(z_{m-1}\beta)]. \quad (17)$$

The appearance of the term $h_{0j} \Delta t_j$ on the righthand sides of (15) and (16) might seem to complicate the formulation. The term represents the increment in the baseline cumulative hazard over the time interval $[t_{j-1}, t_j)$. The term, however, simply reminds us that the covariate vector should include a set of indicator variables to capture the effects of the different stages j of the longitudinal process. In essence, the regression coefficient for the indicator variable for stage j equals $\ln(h_{0j} \Delta t_j) - \ln(h_{01} \Delta t_1)$, i.e., the difference in logarithms of the cumulative baseline hazard increments for the j th and first stage, where it is assumed

that the first stage forms the reference indicator category. Incorporating the term into the covariate vector recasts the baseline hazard function as one that reflects a constant baseline hazard h_{01} (or, equivalently, a baseline exponential survival distribution). If the covariate set is especially rich it may happen that the set of stage indicator variables do not explain a significant amount of variation in the longitudinal sequences and may be discarded from the model.

9. A Case Illustration: Nurses' Health Study of Lung Cancer Risk

The Nurses' Health Study was established in 1976. At the outset of the study, a cohort of 121,700 female registered nurses, aged 30 to 55, completed a questionnaire providing baseline information. The women have completed follow-up questionnaires approximately every two years since enrollment. The endpoint of interest here is a diagnosis of primary lung cancer as confirmed from medical records or death certificates. Subjects are followed until this endpoint or loss to follow-up. For this analysis, we look at occurrence of lung cancer in the cohort from 1986 to 2000. This data set consists of observation sequences on 115,768 women. The sequences contain 748,007 observation intervals and represent 1,577,382 person-years at risk. The endpoint was experienced by 1137 women by the year 2000. For more details about this study, see Bain, Feskanich *et al* [10]. Using these data at a single time point, Lee, Whitmore, and Rosner [5] discussed preliminary results for the TR model and compared the results with those from the Cox proportional hazards model. They did not investigate the longitudinal data. We now present a case illustration to convey the practical value of the Markov TR approach in handling the full longitudinal data set.

The completion times of the periodic questionnaire define the time points t_j . The health process $\{Y(t)\}$ describes the health status with respect to lung cancer and is taken as a latent process, so the y_j are unobservable. For this presentation, two covariates are tracked, namely, the subject's age *age* (covariate $z_{1,j-1}$) and the cumulative pack-years of smoking to date *pkys* (covariate $z_{2,j-1}$). These two covariates define the covariate vector $z_{j-1} = (1, z_{1,j-1}, z_{2,j-1})$ at the start of the time interval $(t_{j-1}, t_j]$, where the unit term is included to provide an intercept. The censoring mechanism is assumed to be independent of the health process. The terms $A_j = (z_j, f_j)$ of the resulting observation sequence for each subject are taken as a realization of a stopped Markov process.

For the Markov decomposition procedure to be valid here, the probability of observation A_j must depend only on the preceding observation A_{j-1} in the sequence. This requirement implies that the covariates (age and cumulative smoking) at time point t_j depend only on their levels at time point t_{j-1} and not on the subject's earlier covariate history. Since age advances deterministically, the Markov property holds trivially. For cumulative smoking, the Markov assumption is plausible but may not be adequate. We adopt the assumption here but recognize that the Markov property would need to be checked in a full investigation. No omnibus test for the Markov property that is required by the decomposition procedure has yet been developed. This issue is a subject for future research.

We use the factorization of $P(A_j|A_{j-1})$ presented in (11). We take the process $\{Y(t)\}$ as a Wiener diffusion process. Each time interval $(t_{j-1}, t_j]$ has two parameters, namely, the initial health level y_{j-1} and the mean parameter μ_{j-1} . The variance parameter is set to 1 because the health status scale is unobservable and, hence, one parameter of the system can be specified arbitrarily. We use a natural logarithmic link function $\ln(y_{j-1}) = z_{j-1}\beta$ for parameter y_{j-1} , where β is a vector of regression coefficients. We set the mean parameter μ_{j-1} to zero for reasons that we discuss later. No distinction is made between alternative time and calendar time for this illustration. The observation sequences of the nurses are assumed to be probabilistically independent.

Each nurse contributes to the sample (partial) likelihood according to her observation sequence. We denote the j th time point for the i th nurse by $t_j^{(i)}$ and the corresponding time interval by $\Delta t_j^{(i)} = t_j^{(i)} - t_{j-1}^{(i)}$, where $j = 1, \dots, m_i$, $i = 1, \dots, n$, and m_i is the length of the observation sequence for nurse i . The cumulative distribution function $F(\cdot)$ and survival function $\bar{F}(\cdot) = 1 - F(\cdot)$ for this case application are those of the inverse Gaussian distribution because this distribution describes the FHT in a Wiener process to a fixed boundary. We denote the parameters of the inverse Gaussian distribution for the i th nurse and j th time interval by $\mu_{j-1}^{(i)}$ and $y_{j-1}^{(i)}$. If the i th nurse survives the j th time interval cancer-free the survival probability $\bar{F}(\Delta t_j^{(i)} | \mu_{j-1}^{(i)}, y_{j-1}^{(i)})$ is contributed to the sample likelihood. If she receives a diagnosis of lung cancer in the j th time interval she contributes the cumulative probability $F(\Delta t_j^{(i)} | \mu_{j-1}^{(i)}, y_{j-1}^{(i)})$. In this case application, we set $\mu_{j-1}^{(i)}$ to zero and use the logarithmic link function $\ln(y_{j-1}) = z_{j-1}\beta$ for the regression function of the initial health-level parameter. Summing over all time intervals for all nurses gives the following log-likelihood function to be maximized.

$$\ln L(\beta) = \sum_{(i,j) \in S} \ln \bar{F}(\Delta t_j^{(i)} | 0, y_{j-1}^{(i)}) + \sum_{(i,j) \in F} \ln F(\Delta t_j^{(i)} | 0, y_{j-1}^{(i)}) \quad (18)$$

Here S and F denote pairs of subscripts (i, j) for intervals and subjects that include the instances of survival and failure, respectively. The cumulative distribution in this special case has the simple form:

$$F(\Delta t_j^{(i)} | 0, y_{j-1}^{(i)}) = 2\Phi \left(-\frac{y_{j-1}^{(i)}}{\sqrt{\Delta t_j^{(i)}}} \right) \quad (19)$$

This cumulative distribution is that of the FHT for Brownian motion because the mean parameter of the Wiener process has been set to zero. We use a numerical gradient method to find the maximum likelihood estimate of vector β .

Table I shows output for our chosen regression model. The model takes account of interval censoring when endpoints lie within observation intervals. The estimates are exact maximum likelihood estimates. The covariates *pkys* and *age* have been centered on their sample mean values of 13 pack-years and 58 years, respectively, to reduce multicollinearity with the intercept coefficient. The regression coefficients of both covariates are significant and negative for the logarithm of the initial health level, $\ln(y_{j-1})$, with P-values < 0.001. The coefficients are roughly of the same order of magnitude. The result implies that the initial health level (with respect to lung cancer) for each observation interval tends to decline with greater age and greater cumulative smoking, indicating a more rapid advance toward the lung cancer endpoint. Smoking an additional pack-year has the same expected decrement in parameter $\ln(y_{j-1})$ as an extra 0.8 years of aging. The sample likelihood function for this data set is insensitive to the value given to the mean parameter μ , which is set to zero. Choosing different values for μ leads to changes in the intercept term for $\ln(y_{j-1})$ but has almost no effect on the regression coefficients for the covariates. The short uniform nature of the observation intervals in this data set (about two years) makes it difficult to separate the effects of $\ln(y_{j-1})$ and μ .

10. Alternative Time Scale and Collapsible Survival Model

The TR model discussed in this article has a connection to models with alternative time scales as found in Cox and Oakes [11], Oakes [12], Kordonsky and Gertsbakh [13], and Duchesne and Lawless [14]. The TR model is also related to collapsible survival models investigated by Duchesne and Rosenthal [15]. In the context of human health, a collapsible survival model postulates that residual survival time is completely determined by the cumulative degradation in the subject's health at that time point. The model invokes the Markov property by assuming that the sample path by which the degradation has accumulated is not relevant; it is only the current level of degradation that determines future survival prospects. The cumulative degradation can be likened to our alternative time scale. Cumulative degradation is a function of calendar time, selected cumulative measures of physical 'wear and tear', and possibly other covariates.

In terms of our notation here, a collapsible model has the following form at a given time point t_j :

$$P(S > t_j | z_j, z_{j-1}, z_{j-2}, \dots, z_0) = \bar{F}[\varphi(t_j, z_j)], \quad \text{for } j=0, 1, 2, \dots, \quad (20)$$

where $\bar{F}(\cdot)$ represents a standard survival distribution function with time origin t_0 . The quantity $\varphi(t_j, z_j)$ is the cumulative degradation of the subject at calendar age t_j , which depends only on age t_j and the covariate vector z_j at that age. By definition, the subject is assumed to be 'new' at age t_0 , so $\varphi(t_0, z_0) = 0$ and $\bar{F}[\varphi(t_0, z_0)] = \bar{F}(0) = 1$. The righthand side of (20) does not depend on prior values of the covariate vectors, i.e., it does not depend on $z_{j-1}, z_{j-2}, \dots, z_0$. It is this latter feature that corresponds exactly to the Markov property of the decomposition procedure. From (20), we can derive the conditional survival probability:

$$P(S > t_j | z_j, z_{j-1}, z_{j-2}, \dots, z_0, S > t_{j-1}) = \frac{\bar{F}[\varphi(t_j, z_j)]}{\bar{F}[\varphi(t_{j-1}, z_{j-1})]}, \quad \text{for } j=1, 2, \dots \quad (21)$$

From the righthand side in (21), it is evident that the conditional probability of the survival time S exceeding t_j , given that it exceeds t_{j-1} , depends only on z_j and z_{j-1} . The observation confirms that the collapsible model embeds an implicit decomposition structure for longitudinal data.

The partial likelihood contribution is obtained by multiplying the terms in (21) over the observation sequence of the subject. The resulting product is the expression on the righthand side of (20), which forms a straight forward basis for inference. Put in simple terms, the only relevant information from the subject up to time point t_j is the subject's covariate vector z_j at that time point. Thus, the unknown parameters of the transformation φ and the standard survival function $\bar{F}$ can be estimated from the longitudinal data using (20) with t_j set to t_m , the stopping time of the study, and z_j set to z_m .

We now demonstrate the collapsible time model using the Nurses' Health Study. The standard survival function $\bar{F}$ is taken to be the FHT distribution for a Wiener diffusion process that starts at health level $\ln(y_0)$ at birth ($t_0 = 0$). The mean parameter μ of the Wiener process is set to zero so the results can be related more closely to the case analysis in Section 9. In effect, therefore, the FHT distribution is that of a Brownian motion process. The standard deviation of the diffusion process is set to unity. The cumulative degradation measure is the extent of disease progression towards the lung cancer endpoint. We assume this measure has the following form: $\varphi(t, z) = t + \beta z$. Here t is the nurse's age in calendar

years and z is her cumulative amount of smoking to that age, measured in pack-years and denoted later by $pkys$. The age variable t defines the measurement unit of the alternative time scale and, hence, has a regression coefficient equal to one. Thus, the alternative time scale φ is measured in effective years of age.

The nurses entered this study at different ages so the collapsible model must take account of the left-truncation that is implicit in this staggered recruitment. Thus, for a nurse who enters the study at age t_1 with cumulative smoking to date of z_1 and has a stopping time of t_m , the likelihood contribution is the conditional probability $P(A_m|A_1)$. Specifically, if the nurse is surviving at the end of follow-up t_m with cumulative smoking to date of z_m then her likelihood contribution is the following conditional survival probability derived from (20):

$$P(A_m|A_1) = \frac{\bar{F}[\varphi(t_m, z_m)]}{\bar{F}[\varphi(t_1, z_1)]}. \quad (22)$$

If the subject has a diagnosis of lung cancer at time t_m then (22) is replaced by the corresponding conditional probability density function.

Table II gives the maximum likelihood estimates for parameters of the collapsible time model in (22). Parameter $\ln(\beta)$ differs significantly from zero with a P-value < 0.001 . The regression coefficient for $\ln(\beta)$ is shown as -0.3730 , which corresponds to an estimate for β of $\exp(-0.3730) = 0.69$. This result suggests that each pack-year of smoking is equivalent to adding seven-tenths of a year to the smoker's age. Put another way, a nurse who will smoke one pack (per day) every year from age 20 to age 50 and then quits smoking will have the same lung cancer risk as a non-smoker at age 71. This result for the effect of smoking on lung cancer risk roughly matches the result found in the Markov TR case analysis in Section 9, although the models are different in key aspects.

11. Concluding Review and Discussion

The threshold regression (TR) model encompasses a wide class of survival models that are important in medical applications. The model is based on the simple idea that a medical endpoint occurs when the underlying health process reaches a boundary for the first time. Most survival studies also involve time-varying covariates. When health status and covariate readings for subjects are gathered at a sequence of time points, the data set includes longitudinal observation sequences that terminate in either censoring or failure. The formulation allows observations to be unevenly spaced in time. This article presents an approach to these longitudinal data called *threshold regression with Markov decomposition*; abbreviated *Markov TR*. The approach is based on the plausible assumption that the observation sequence for each subject is a stopped Markov process. This assumption allows observation sequences to be decomposed into a series of single records that can be handled and analyzed with the same ease as cross-sectional data. The article explores a number of important theoretical and practical aspects of the approach and demonstrates its application with a real case study. As stressed in earlier publications on TR, the model encourages investigators to look carefully at the underlying stochastic mechanism by which the health process evolves and an endpoint is triggered. The mathematical forms of the stochastic process, boundary and time scale are all elements that play an important part in describing and understanding this mechanism. The extension to the longitudinal data context not only eases application of the TR model but also expands understanding and insights that investigators gain from its application.

The article leaves a number of interesting and important research questions unresolved and in need of further study. Aside from choosing the right stochastic process, boundary and time scale, the best choice of covarying processes is an important research issue. Equally important is knowing the appropriate form of regression functions and link functions connecting the covarying process to the health process. Methods of model checking and validating the Markov property and, hence, the propriety of the proposed decomposition procedure requires additional theoretical and practical research. The Cox proportional hazards regression model (PH model), with and without time-covarying covariates, has been a reliable standby for medical survival data analysis. The reaction of analysts to this proposed Markov TR approach may very well be summed up by the question, ‘The PH model works well for me, so why is another approach needed?’ Yet, on some reflection, the analyst may see that the PH model provides hazard ratios but goes no deeper. Markov TR encourages construction of a richer model - a model that can offer a more profound interpretation of results. It can help an investigator to formulate a causal explanation for the endpoint by pointing to the sources and magnitudes of risk emanating from the underlying health process, boundary or time scale. For example, it might show that a lower hazard rate associated with a treatment is a consequence of an immediate improvement in the health level at baseline and not from a slower rate of disease progression. Also, where considered at the planning stage of a study, the Markov TR model offers a more thoughtful approach to the choice of study design, sample size, data analysis and inference methods. Many researchers have seen the value of jointly modeling longitudinal measurements and event-time data. Henderson *et al* [16], for example, propose a bivariate Gaussian process (possibly correlated) in which one component drives a longitudinal measurement process and the other drives an intensity process for events. The bivariate process is taken to be a latent process. The purpose of their approach, like ours, is to accommodate longitudinal and event data together. The generalized time series model proposed by Zeger and Qaqish [17] is another example of a model that might be expanded by melding it with ideas drawn from the Markov TR model. Finally, the exploration of Markov TR applications in new areas of medicine also offers a wide and potentially fruitful research agenda.

Acknowledgments

This research is supported in part by NIH Grants OH008649 (Lee), EY012269 (Rosner), HL40619 (Rosner) and by a research grant from the Natural Sciences and Engineering Research Council of Canada (Whitmore).

References

1. Lee M-LT, Whitmore GA. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science*. 2006; 21:501–513.
2. D’Agostino RB, Lee M-LT, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine*. 1990; 9:1501–1515. [PubMed: 2281238]
3. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*. 1988; 83:414–425.
4. Aalen, OO.; Borgan, A.; Gjessing, HK. *Survival and Event History Analysis: A Process Point of View*. Springer; New York: 2008.
5. Lee, M-LT.; Whitmore, GA.; Rosner, BA. Benefits of Threshold Regression: A Case-study Comparison with Cox Proportional Hazards Regression. 2009. submitted, under review
6. Lee M-LT, DeGruttola V, Schoenfeld D. A model for markers and latent health status. *J Royal Statist Soc Series B*. 2000; 62:747–762.
7. Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2. John Wiley & Sons; New York: 2002.
8. Hougaard, P. *Analysis of Multivariate Survival Data*. Springer; New York: 2000.

9. Whitmore GA, Crowder MJ, Lawless JF. Failure inference from a marker process based on a bivariate Wiener model. *Lifetime Data Analysis*. 1998; 4:229–251. [PubMed: 9787604]
10. Bain C, Feskanich D, Speizer FE, Thun M, Hertzmark E, Rosner BA, Colditz GA. Lung cancer rates in men and women with comparable histories of smoking. *Journal of the National Cancer Institute*. 2004 June 2.96(11)
11. Cox, DR.; Oakes, D. *Analysis of Survival Data*. Chapman and Hall; London: 1984.
12. Oakes D. Multiple time scales in survival analysis. *Lifetime Data Analysis*. 1995; 1:7–18. [PubMed: 9385093]
13. Kordonsky KB, Gertsbakh I. Multiple time scales and the lifetime coefficient of variation: engineering applications. *Lifetime Data Analysis*. 1997; 3:139–156. [PubMed: 9384619]
14. Duchesne T, Lawless J. Alternative time scales and failure time models. *Lifetime Data Analysis*. 2000; 6:157–179. [PubMed: 10851840]
15. Duchesne T, Rosenthal JS. On the collapsibility of lifetime regression models. *Advances in Applied Probability*. 2003; 35:755–772.
16. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event-time data. *Biostatistics*. 2000; 1:465–480. [PubMed: 12933568]
17. Zeger SL, Qaqish B. Markov regression models for time series: A quasi-likelihood approach. *Biometrics*. 1988; 44:1019–1031. [PubMed: 3148334]

Table I

Results for a threshold regression model with Markov decomposition of observation sequences in the Nurses' Health Study. The initial log-health level $\ln(y_{j-1})$ for each observation interval depends on initial cumulative smoking and initial age, denoted by $pkys$ and age . Parameter μ is set to zero. The covariates are centered on their sample mean values of 13 pack-years and 58 years of age, respectively.

Parameter	Variable	Estimate	Std. Error	P-value
$\ln(y_{j-1})$				
	pkys - 13	-.0053	.0001	< 0.001
	age - 58	-.0066	.0004	< 0.001
	intercept	1.6032	.0039	

Table II

Fitted collapsible survival model in which the time scale is a linear combination of age, measured in calendar years, and cumulative smoking (*pkyrs*), measured in pack-years.

Parameter	Variable	Estimate	Std. Error	P-value
$\ln(y_0)$				
	intercept	3.0628	.0081	
$\ln(\beta)$				
	pkyrs	-.3730	.0740	< 0.001