Independent Study of : "A model for markers and latent health status by Lee et al

Kiana Gravesande 001301267

09/12/2021

Summary

Lee et al, use a bivariate Wiener diffusion process as the underlying stochastic process of a threshold regression model to model the joint process of a marker and health status. Lee et al extend the bivariate Wiener process presented by Whitmore et al.[1] In this model, the health status is latent. The time to failure, i.e death or onset of disease, is the time at which the health status first reaches a boundary or threshold. Unlike the cox regression model, threshold regression does not rely on the proportional hazards assumption and can be applied in a variety of situations. The bivariate threshold regression model was applied to AIDS data because antiviral resistance lowers the effects of antiretroviral drugs over time; therefore, the proportional hazards assumption may not be reasonable in clinical trials for AIDS treatments. CD4 cell count is used as a marker process that tracks the patient's progression to death.[2]

Introduction

Threshold Regression refers to first hitting time (FHT) models with regression structures that relate the model parameters to covariates. Lifetime or time-to-event data may be interpreted as the first passage time to a failure threshold or boundary state of a stochastic process (i.e first hitting time). In a clinical context, this failure threshold may be death or disease progression. First hitting times occur in a variety of stochastic models, such as Wiener, Poisson, Bernoulli and Gamma processes.[3] FHT models can be applied to many fields such as environmental science, medicine, economics and engineering. In engineering, FHT models can be used to model the time to failure of a mechanical system.[1] In economics, Lancaster uses a Wiener FHT model to model the length of a strike.[4] Lee et al extend a Biviarte Wiener Model by Whitmore et al. Whitmore et al present a model for the failure of an engineering mechanical system. In this model the process of degradation is assumed to be latent. Whitmore et al include an observable marker process which is time-varing and tracks the mechanical system's degradation. Whitmore et. al do not incorporate covariates in their model. Lee et al. extend this model to incorporate covariates and apply their model to a clinical trial of AIDS patients to show the usefulness of this model.

Methods

First Hitting Time Models

Using notation from Lee and Whitmore[3], a FHT model has two main components: i. A Parent Stochastic Process $\{X(t), t \in \mathbb{T}, x \in \mathbb{X}\}$ with an initial value $X(0) = x_0$, where $\mathbb{X}$ is the state space and $\mathbb{T}$ is the time space of the process and ii. A boundary set $\mathbb{B}$, where $\mathbb{B} \subset \mathbb{X}$. The initial value $X(0) = x_0$ of the process is outside of the boundary set $\mathbb{B}$ and the FHT of $\mathbb{B}$ is the random variable S where $S = \inf\{t : X(t) \in \mathbb{B}\}$. The process $\{X(t)\}$ may have a different properties, such as the Markov property or monotonic sample paths. The sample path of the parent process, $\{X(t)\}$, can be observable or latent. Latent processes are the most commonly used. In some FHT models, the process $\{X(t)\}$ is not guaranteed to

encounter the boundary set, i.e $P(S < \infty) < 1$. A scenario where an infinite FHT is possible is in a Wiener diffusion model with a boundary at zero and a drift away from the boundary, i.e $\mu > 0$. A finite FHT is not guaranteed in this scenario.[2]

Data Model for Threshold Regression

The data structure of threshold regression varies. Taking an example where longitudinal observations are available for the parent process and the covariate vector, below are the general elements of the data structure of an individual, i ,[3,5]:

(1) Time points: $0 = t_0 \leq t_1 \leq \cdot \cdot \cdot \leq t_m$,
(2) Failure Indicator: $f_0 = 0, f_1 = 0, ..., f_{m-1} = 0, f_m = 0 \, or \, 1$
(3) Measurements of parent process: $x_0, x_1, ..., x_m$
(4) Covariate vector: $z_0, z_1, \ldots, z_m$. Regression link functions connect the covariate vector to parameters. The link function will have the general form $\theta_j = g(z_j \beta_\theta)$, j = 1,. . . ,n, for each parameter θ

If m = 1, i.e there is only one time point and failure indicator available for each individual, then the data is censored survival data.

A bivariate model for markers and latent health status

Lee et al construct a bivariate model for markers and latent health status. One component of this bivariate model is {X(t)} which is a latent process that represents the health status of an individual at time t. The progression of disease corresponds to a decline in health status. Let $\{Y_w\}$ denote an observed marker process which is correlated with the health status process and helps to track its progress. Lee et al introduce a two-dimensional Wiener diffusion process $\{X(t), Y_w(t)\}$, for t greater than or equal to zero and having initial values $\{X(0), Y_w(0)\}$. This two-dimensional Wiener Process has a bivariate normal distribution which has mean vector $\{X(0), Y_w(0)\} + t\mu$, where $\mu = (\mu_x, \mu_y)$, and covariance matrix $t\Sigma$, where

$$\Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix}.$$

The parameter, μ_x, is the mean change per unit time in an individual's health status. If μ_x is negative, this means the individual's health status is decreasing and tends to drift towards the failure threshold or boundary. The parameter, μ_y, is the mean change per unit time in the marker process. The parameter, σ_{xx}, is the inherent variability per time in health status and the parameter σ_{yy} is the inherent variability per time in the marker process. The parameters, σ_{xy} and σ_{xy}, are the covariance of the health status and marker process. The individual's initial health status is denoted by $\delta = X(0) > 0$, where δ is not known and is estimated. Lee et al define the failure threshold to be 0. The nearer X(t) is to 0, the sicker the individual is. The point of failure is the first time the individual's health status decreases to zero. Lee et al adjust the marker process to represent changes in the marker level from the initial value. This marker change process is $\{Y(t)\} = Y_w(t) - Y_w(0)$. Following the earlier notation, the first passage-time from the initial health status to the failure threshold

or boundary is denoted by the random variable S.[2] How well the marker process tracks the parent process depends on how strongly correlated the marker process is with health status and it's inherent variability. Examples of marker processes are blood pressure for cardiovascular disease and CD4 cell count for AIDS.[3] The correlation of the health status and marker process is

$$\rho = \sigma_{xy}/(\sigma_{xx}\sigma_{yy})^{1/2}$$

Using notation from Lee et al[2], Suppose that each individual is observed for a time period (0,t] and has two possible observation outcomes:

1. The individual survives to time t at which time a marker value of Y(t) = y(t) is documented. This is a censored observation of survival time because S > t
2. The individual fails at some time S = s, which lies in (0,t], and a marker value of Y(S) = y(s) is documented at the failure time.

The conditional density function of Y(t)|x(t) for an individual who survives to time t has the following form:

$$p_1(y|x) = \lim_{h\to 0}\left[\frac{1}{h}Pr\{Y(t) \in (y, y+h)|X(t) = x, S > t\}\right]$$

The conditional density function Y(s)|x(s) for an individual who fails at time s, where S = s and S $\leqslant$ t is

$$p_2(y|s) = \lim_{h\to 0}\left[\frac{1}{h}Pr\{Y(s) \in (y, y+h)|S = s\}\right].$$

The above densities, p_1 and p_2, are the distributions of conditional normal variables N$\{\mu_{y.x}$(r),$\sigma_{yy.x}$r$\}$, where r = s for a failing individual and r = t for a survivor. The conditional mean and variance are

$$\mu_{y.x}(r) = \mu_y r + \sigma_{xy}\sigma_{yy}^{-1}\{x(r) - \delta - \mu_x r\},$$

$$\sigma_{yy.x} = \sigma_{yy}(1-\rho^2).$$

Below is the density for an individual who survives past time t and has health status X(t) at time t:

$$p_3(x) = \lim_{h\to 0}\left[\frac{1}{h}Pr\{X(t) \in (x, x+h), S > t\}\right] = (2\pi\sigma_{xx}t)^{-1/2}exp\left\{\frac{(x-\delta-\mu_x t)^2}{2\sigma_{xx}t}\right\}f(x),$$

where 0 < x < ∞ and f(x) = 1 - exp(-2δx/$\sigma_{xx}t$).

The survival time S, of an individual who fails, has an Inverse Gaussian (IG) distribution with density function:

$$p_4(s) = \lim_{h\to 0}\left[\frac{1}{h}Pr\{S \in (s, s+h)\}\right] = \delta(2\pi\sigma_{xx}s^3)^{-1/2}exp\left\{-\frac{(\delta+\mu_x s)^2}{2\sigma_{xx}s}\right\}.$$

The density functions for the two aforementioned outcomes can be derived as follows the from the results presented above:

1.) The joint density for an individual surviving past time t and having health status X(t) and marker level Y(t) at time t is $p_1(y|x)p_3(x)$. As x is latent, x can be integrated out of this joint density:

$$p_s(y) = \lim_{h\to 0}\left[\frac{1}{h}Pr\{Y(t) \in (y, y+h), S > t\}\right] = \int_0^\infty p_1(y|x)p_3(x)dx$$

2.) For an individual who fails at time s, the joint density of Y(s) and S is:

$$p_d(y,s) = \lim_{h_1\to 0, h_2\to 0}\left[\frac{1}{h_1 h_2}Pr\{Y(s) \in (y, y+h_1), S \in (s, s+h_2)\}\right] = p_2(y|s)p_4(s)$$

If the individual fails or not is Bernoulli random variable with probability P_d where

$$P_d = P(S \leqslant t) = \int_0^t \int_{-\infty}^\infty p_d(y,s)dyds$$

For a sample with n individuals observed during the time interval (0,t], the sample observations on failing individuals is denoted by $(y_i,\ s_i)$, i= 1,. . . , k where k is the number of failures. The sample observations on the survivors is denoted by y_i, i = k+1,. . . ,n. The log-likelihood of the sample is[2]:

$$log\{L(\mu,\Sigma)\} = \sum_{i=1}^{k} log\{p_d(y_i,s_i)\} + \sum_{i=k+1}^{n} log\{p_s(y_i)\}$$

.

$log\{L(\mu,\Sigma)\}$ depends only on μ_x/δ, μ_y, ρ, σ_{yy} and σ_{xx}/δ^2. Consequently, one of the parameters of the latent health status process can be fixed. The authors set σ_{xx} to one since {X(t)} is latent.

Lee et al derive the density function to predict the health status of an individual at time t from their current marker value. Using the notation from earlier, this density function is:

$$p_5(x|y) = \lim_{h\to 0}\left[\frac{1}{h}\Pr\{X(t) \in (x,\ x+h) | Y(t) = y,\ S > t\}\right] = \frac{p_1(y|x)\,p_3(x)}{p_s(y)}, \qquad 0 < x < \infty.$$

Lee et. al also present the density function to predict the residual life of a survivor given the individual's current marker value. This conditional density is:

$$p_7(s|y) = \lim_{h\to 0}\left[\frac{1}{h}\Pr\{S \in (s,\, s+h)|\, Y(t) = y,\, S > t\}\right] = \int_0^\infty p_6(s|x)\, p_5(x|y)\, \mathrm{d}x, \qquad t < s < \infty.$$

where $p_6(s|x)$ is the failure time distribution of a surviving individual with health status x at time t which is Inverse Gaussian, where s is replaced by s - t and δ, the initial health status, is replaced by x.

These inferences are made after computing the maximum likelihood estimates of the model parameters. The data should follow an Inverse Gaussian (IG) distribution, the first stopping time distribution of a Wiener process. The Inverse Gaussian survival curves can be compared against the Kaplan-Meier Survival Curve[2,5].

Covariates And Link Functions

Covariates are incorporated into the bivariate Wiener FHT model. The parameters of the bivariate Wiener FHT model are assumed to be connected to a linear combination of covariates by a suitable link function as follows:

$$\begin{aligned}
\delta &= g_\delta(\mathbf{Z}\beta_\delta),\\
\mu_x &= g_x(\mathbf{Z}\beta_x),\\
\mu_y &= g_y(\mathbf{Z}\beta_y),\\
\sigma_{yy} &= g_{yy}(\mathbf{Z}\beta_{yy}),\\
\rho &= g_p(\mathbf{Z}\beta_\rho),
\end{aligned}$$

where $\mathbf{Z}$ is the covariate vector and β_i, for each parameter i, is a vector of regression coefficients. The covariates used may be fixed in time, such as age at infection or time-varying,e.g medication rate which is increased or decreased over time[2]. The suitability of the link function used should be assessed.

The authors suggest identity links for μ_x and μ_y, i.e. $\mu_x = \mathbf{Z}\beta_x$ and $\mu_y = \mathbf{Z}\beta_y$. Since δ and σ_{xx} are always positive, the authors suggest $\delta = \exp(\mathbf{Z}\beta_\delta)$ and σ_{yy} =$\exp(\mathbf{Z}\beta_{\sigma_{yy}})$. Since ρ is in the interval (-1,1) the following link function is suggested:

$$\rho = \frac{exp(\mathbf{Z}\beta_\rho) - 1}{exp(\mathbf{Z}\beta_\rho) + 1}.$$

Data and Application of Model

Lee et al apply the bivariate model for markers and latent health status to data from the AIDS Clinical Trial Group 116a study. The number of individuals included in this application of the model was 787. Lee et al use death as the failure point. The Marker process {Y(t)} in this application is the natural logarithm of the ratio of the final CD4 count before follow-up or death and the baseline CD4 count. An indicator variable, fail, denotes whether or not the individual dies during the time period of the trial. The variable, time, represents the number

of days from the start of treatment to death or to the end of follow-up; log(cd0) is the natural logarithm of the baseline CD4 cell count for a patient; d500 and d750 are indicator variables indicating the administration of 500 mg per day and 750 mg per day of didanosine, where zidovudine is the standard (reference) treatment.The variable, prev, represents the length in years of previous zidovudine treatment; d500prev and d750prev are interaction variables calculated from the products of d500 and d750 with prev.

Model Assumptions

The Wiener process assumes independence of increments and proportionality of the mean and variance parameters to the time increment. These assumptions can be checked for the marker process. This was done by plotting the standardized residuals w_{ij} against time increments Δt_{ij}. The results are shown in Fig. 1. There is no trend in the residuals as a function of the time increment. This confirms that the mean and variance of the marker increments are proportional to the time increment. A normal probability plot of the standardized residuals is also inspected to determine if the quantities are normally distributed per the assumption of a Wiener marker process. The results are shown in Fig 2. From figure 2, this assumption holds. The outliers seen are either very small initial or terminal CD4 cell count. Small CD4 cell counts have high variability; therefore, these outliers are expected.[2]

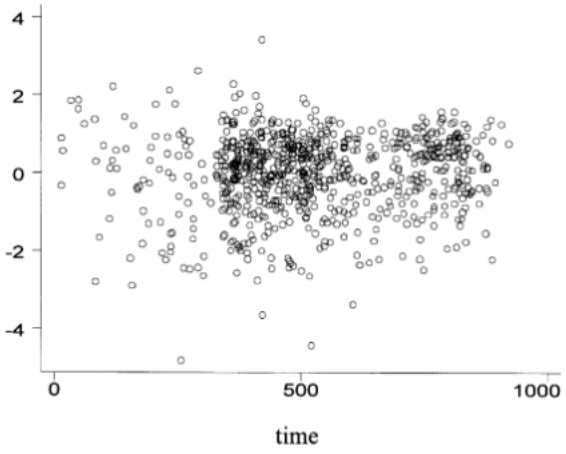

Fig. 1. Residual plot

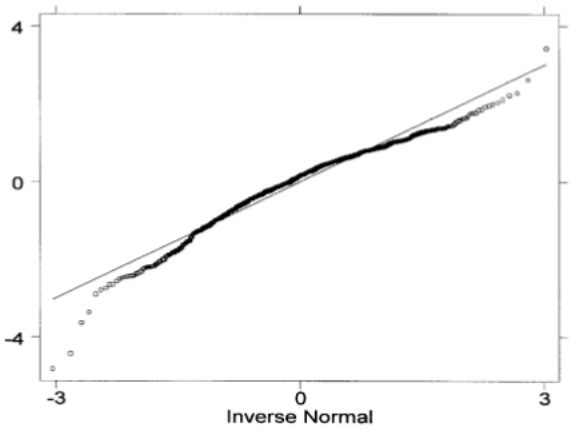

Fig. 2. Normal probability plot

Results

The results of the application of the bivariate model for markers and latent health status, referred to as the Marker Model, are shown below in Table 1. $\log(\delta)$, where δ is the initial health status, is significantly associated with log(cd0), the natural logarithm of the baseline CD4 cell count for a patient. The regression coefficient for log(cd0) is 0.270. This means that individuals who enter the study with lower baseline CD4 count are nearer to failure (i.e death). For parameter μ_y, previous use the standard regimen, prev, has a significant effect on the marker process. The regression coefficient for prev for the parameter μ_y is -0.00500. This negative value confirms that previous use of zidovudine results in a faster decline to death. The correlation between the health status process and marker process was estimated to be 0.242. Since the CD4 marker and health status are only moderately correlated, use of the CD4 cell count to track the progress of AIDS may not very valuable.[2]

Table 1. Model 1 (the marker model): marker and latent parameter estimation†

Marker model parameter	*Regression covariate*	*Estimated coefficient*	*Standard error*	*z*	*P > |z|*
$\log(\delta)$	log(cd0)	0.2704	0.0275	9.827	< 0.001
	constant	2.4637	0.1312	18.780	< 0.001
μ_x	prev	−0.0956	0.0424	−2.254	0.024
	d500	−0.0136	0.0069	−1.971	0.049
	d750	−0.0057	0.0068	−0.834	0.404
	d500prev	0.1242	0.0694	1.790	0.074
	d750prev	0.1293	0.0609	2.123	0.034
	constant	−0.0121	0.0059	−2.033	0.042
μ_y	prev	−0.00500	0.0013	−3.898	< 0.001
	d500	0.00028	0.0002	1.434	0.152
	d750	−0.00002	0.0002	−0.082	0.935
	d500prev	−0.00091	0.0020	−0.463	0.644
	d750prev	0.00254	0.0017	1.456	0.145
	constant	−0.00103	0.0001	−8.338	< 0.001
$\log(\sigma_{yy})$	constant	−6.3165	0.0501	−146.126	< 0.001
$\log\{(1+\rho)/(1-\rho)\}$	constant	0.4930	0.0882	5.588	< 0.001

†Log-likelihood 350.32.

The authors compare the results of this threshold regression model to Kaplan-Meier (KM) survivor function estimates. Since the KM survivor function cannot accommodate the marker data, the authors use the generalized model of Whitmore et al, which does not take into account "post-base-line marker response"[2]. The baseline CD4 values are used in this generalized model. This model is no longer a bivariate model and is now a censored IG regression model, where the survival function is constructed from the IG distribution. The results of this model are shown in Table 2. These results agree with the results of Model one. Specifically, $\log(\delta)$,is shown to be significantly associated with log(cd0).

Table 2. Model 2: a censored IG regression without marker information†

Model 2 parameter	*Regression covariate*	*Estimated coefficient*	*Standard error*	*z*	*P > |z|*
$\log(\delta)$	log(cd0)	0.2660	0.0281	9.459	< 0.001
	constant	2.4810	0.1337	18.554	< 0.001
μ_x	prev	−0.0974	0.0424	−2.297	0.022
	d500	−0.0144	0.0069	−2.073	0.038
	d750	−0.0061	0.0069	−0.896	0.370
	d500prev	0.1249	0.0690	1.810	0.070
	d750prev	0.1326	0.0607	2.183	0.029
	constant	−0.0118	0.0060	−1.968	0.049

†Log-likelihood 222.12.

Figure 1: Patient Characteritics

In comparing the KM survivor function estimates to the censored IG estimates, the authors examine the data over a range of baseline of values of CD4 cell count for the IG model. The survival function for an individual who survived to at least week 16 of the trial and who received the standard treatment, zidovudine, and who did not previously use zidovudine was calculated from the Inverse Gaussian Model. Five IG survival curves which correspond to the 10th, 25th, 50th, 75th and 90th percentiles of baseline CD4 cell counts were generated. These five curves and the Kaplan Meier curve are shown below in Fig 3. The KM residual survival curve includes all individuals having any initial CD4 cell count, who survived to at least week 16 of the trial and who received the standard treatment, zidovudine, and who did not previously use zidovudine.

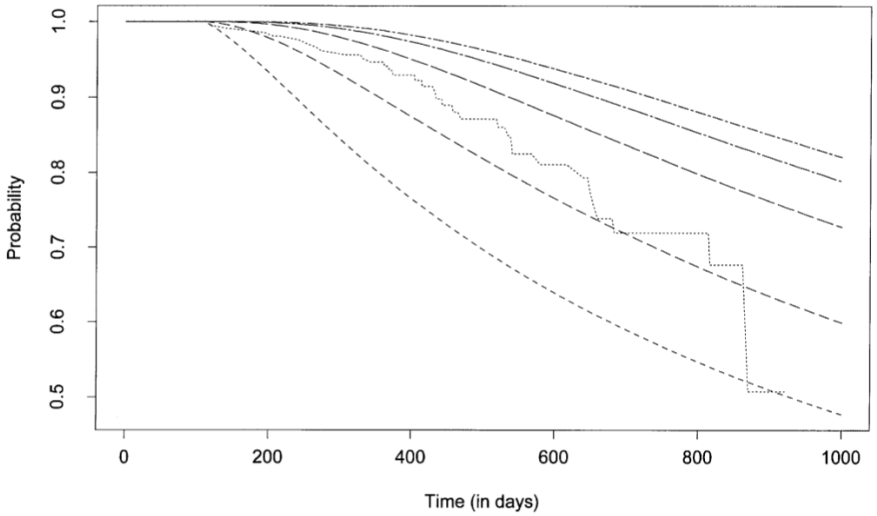

Fig. 3. Predicted and observed AIDS-free survival for subjects treated with zidovudine: - - - - -, 10th percentile; – – – –, 25th percentile — —, 50th percentile; · — ·, 75th percentile; · – · –, 90th percentile ·····, KM curve

The KM survival curve runs through the middle of the five IG survival curves, as expected. The average of these five IG survival curves closely approximates the KM curve. From the KM curve, the estimate of probability of survival at day 600 is 0.81. The average of estimated probabilities of survival estimated from the five IG curves is also approximately 0.81 at 600 days. This shows that the IG model fits the data well.[2]

Conclusion

Threshold regression models are useful in scenarios where the proportional hazards assumption does not hold. An example of such scenario is a clinical trial where the rate of medication is increased or decreased over time. The suitability of the FHT model used should be assessed. Input from subject matter experts may be needed to determine which FHT models may be suitable.

References

1. Whitmore GA, Crowder MJ, Lawless JF. Failure inference from a marker process based on a bivariate Wiener model. Lifetime Data Anal. 1998;4(3):229-251. doi:10.1023/a:1009617814586

2. Ting Lee, M., DeGruttola, V. and Schoenfeld, D., 2000. A model for markers and latent health status. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(4), pp.747-762.

3. Lee M-LT, Whitmore GA. Threshold regression for Survival Analysis: Modeling Event Times by a stochastic process reaching a boundary. Statistical Science. 2006;21(4). doi:10.1214/088342306000000330.

4. Lancaster T. A Stochastic Model for the Duration of a Strike. Journal of the Royal Statistical Society Series A (General). 1972;135(2):257. doi:10.2307/2344321

5. Balakrishnan, N. and Rao, C., 2004. Advances in survival analysis. Amsterdam: Elsevier North-Holland.