MODELING TRANSITION PROBABILITIES AND DURATION TIME FOR EVENT HISTORY DATA

Yamashita, Natsumi
Department of Medical Informatics, Kyushu University Hospital

Nagata, Eiko
Biostatistics Center, Kurume University

Ohyama, Tetsuji
General Clinical Research Center, Oita University Hospital

Yanagawa, Takashi
Biostatistics Center, Kurume University

https://doi.org/10.5109/25907

出版情報：Bulletin of informatics and cybernetics. 42, pp.71-79, 2010-12. Research Association of Statistical Sciences
バージョン：
権利関係：

MODELING TRANSITION PROBABILITIES AND DURATION TIME FOR EVENT HISTORY DATA

by

Natsumi Yamashita, **Eiko Nagata**, **Tetsuji Ohyama**
and
Takashi Yanagawa

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.42

FUKUOKA, JAPAN
2010

Bulletin of Informatics and Cybernetics, Vol. 42, 2010

MODELING TRANSITION PROBABILITIES AND DURATION TIME FOR EVENT HISTORY DATA

By

Natsumi Yamashita[*] **Eiko Nagata**[†] **Tetsuji Ohyama**[‡]
and
Takashi Yanagawa[§]

Abstract

There could be covariates that are related to transition probabilities, but not related to duration times to the next event, or vice versa, in event history analyses. A method is proposed in this paper that estimates those relationships, separately, by using an accelerated failure time model. The method is applied to the data taken from individuals staying at home and having government care service, where different levels of care are provided to care need individuals based on their severities officially recognized. It is shown among others that family and disease are significantly related to duration times to the next event, but not related to transition probabilities.

Key Words and Phrases: accelerated failure time model; event history analysis; Markov process; survival analysis

1. Introduction

Let $\{X_n, n = 0, 1, 2, \ldots\}$ be a stochastic process assuming values in the finite set $\mathcal{S} = \{0, 1, 2, \ldots, J+1\}$. Let $T_0, T_1, T_2, \ldots$ be the transition time on the nonnegative half of the real line such that $0 = T_0 < T_1 < T_2 \cdots$. The two dimensional process $(X, T) = \{(X_n, T_n);\ n = 0, 1, 2, \ldots\ \}$ is said to have the semi-Markov property if the condition

$$\begin{aligned} P\{X_{n+1} = j, T_{n+1} - T_n \le t \quad &| \quad (X_k, T_k), k = 0, 1, \ldots, n\} \\ &= \quad P\{X_{n+1} = j, T_{n+1} - T_n \le t \mid X_n\} \qquad (1.1) \end{aligned}$$

is satisfied for all $n = 0, 1, 2, \ldots$ and $t > 0$. Note that the time homogeneous is assumed in this definition in the sense that the right hand side of (1.1) does not depend on n.

Analysis of repeated multiple events, mathematically represented by (X, T), is often called the event history analysis. It is concerned with the (possible repeated)

[*] Biostatistics Center, Kurume University, 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan. Department of Medical Informatics, Kyushu University Hospital, 3-1-1 Maidashi, Higashi-ku, Fukuoka-city, Fukuoka 812-8582, Japan.

[†] Biostatistics Center, Kurume University, 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan.

[‡] General Clinical Research Center, Oita University Hospital, 1-1 Idaigaoka, Hasama-machi, Yuhu-city, Oita 879-5593, Japan.

[§] Biostatistics Center, Kurume University, 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan.

occurrence and duration to next event and, typically, their dependence on some explanatory variables (covariates) (Huang and Stone (1998)). The semi-Markov property is assumed in conventional event history analysis. It enables one to apply the Cox regression by treating the data that transit to the other states as censoring. Beautiful reviews of the event history analysis are given by Andersen and Keiding (2002) and Andersen and Perme (2008), and the method of analysis is illustrated recently by Putter et al. (2007).

The relationship between covariates and duration times to the next event, that has been the focus in conventional analyses, is modeled by means of intensity functions. Transition probabilities may be also computed as functions of the intensity, but it results in to use the same covariates as those that were selected in illustrating the relationship with duration times to the next event. However, as is shown in the application in this paper, there could be a situation where some covariates are related to the transition probability only, or vice versa. If this is the case, it might be better to model relationships with duration times to the next event and transition probabilities, separately.

In this paper we extend the semi-Markov property by representing

$$\begin{aligned} P\{X_{n+1}=j,T_{n+1} &- T_n \le t \mid (X_k,T_k), k=0,1,\ldots,n\} \\ &= P\{T_{n+1}-T_n \le t \mid X_{n+1}=j,(X_k,T_k),k=0,1,\ldots,n\} \\ &\qquad \times P\{X_{n+1}=j \mid (X_k,T_k), k=0,1,\ldots,n\}, \quad (1.2)\end{aligned}$$

and assuming (A) and (B) below regarding the first and second probabilities of the right hand side of equation (1.2).

(A) $P\{T_{n+1}-T_n \le t \mid X_{n+1}=j,(X_k,T_k),k=0,1,\ldots,n\}$
$= P\{T_{n+1}-T_n \le t \mid X_{n+1}=j,(X_n,T_n)\}$ for any $j \in \mathcal{S}$ and $t \ge 0$

and

(B) $P\{X_{n+1}=j \mid (X_k,T_k),k=0,1,\ldots,n\}$
$= P\{X_{n+1}=j \mid (X_n,T_n)\}$ for any $j \in \mathcal{S}$ and $t \ge 0$.

The set of assumptions (A) and (B) is equivalent to assume that

$$\begin{aligned} P\{X_{n+1}=j,T_{n+1}-T_n \le t \mid (X_k,T_k),k=0,1,\ldots,n\} \\ = P\{X_{n+1}=j,T_{n+1}-T_n \le t \mid (X_n,T_n)\}.\end{aligned}$$

The difference of equation (1.1) and this equation is the addition of T_n to the conditional part of the right hand side of equation (1.1). It is trivial that if $P\{T_{n+1}-T_n \le t \mid X_{n+1} = j, X_n, T_n\}$ and $P\{X_{n+1} = j \mid X_n, T_n\}$ do not depend on T_n, then the semi-Markov property holds.

Let z be the p-dimensional covariate vector of an individual. We introduce mathematical models that relate z to the probabilities of the right hand side of (A) and (B), separately, and develop a method of assessing the influence of z on these probabilities. Note that, in conventional methods, functions of z are introduced into the intensity function defined by

$$\lambda_{ij}(t|z) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\{T_{n+1}-T_n \le t+\Delta t, X_{n+1}=j|T_{n+1}-T_n > t, X_n = i, z\}.$$

Foucher et al. (2007) used the same decomposition (1.2) as ours for modeling interval-censored data with multiple terminal events, but they studied intensity functions by assuming the semi-Markov property.

The method developed in this paper will be applied to the data taken from people staying at home and having government care service, where the duration times to next event are times to the deterioration or improvement that officially recognized. It is shown among others that family and disease are significantly related to duration times to the next event, but not related to transition probabilities.

2. Mathematical development

2.1. Assumptions

Let $\{X_n, n = 0, 1, 2, \ldots\}$ be a stochastic process assuming values in the finite set $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$, where $\mathcal{S}_0 = \{0, J+1\}$ is absorbing state and $\mathcal{S}_1 = \{1, 2, \ldots, J\}$ is transition state. Thus $P_{0j} = 0$, $P_{J+1,j} = 0$ for any $j \in \mathcal{S}_1$, and $P_{ij} > 0$ for any $i \in \mathcal{S}_1$ and some $j \in \mathcal{S}$, where P_{ij} is the transition probability from state i to state j. Let $\{T_n, n = 0, 1, 2, \ldots\}$ be the transition times defined in the previous section and z be the covariate vector.

The probability density function of the duration time to the next event conditioned on $X_{n+1} = w$, $X_n = i$, $T_n = s$ and z ($i \in \mathcal{S}_1$, $w \in \mathcal{S}$) is given by

$$\begin{aligned}&f(t|X_{n+1} = w, X_n = i, T_n = s, z)\\&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\{t < T_{n+1} - T_n \leq t + \Delta t \;|X_{n+1} = w, X_n = i, T_n = s, z\}.\end{aligned}$$

Putting $u_{iwsz} = \{X_{n+1} = w, X_n = i, T_n = s, z\}$ for simplicity, we denote the conditional survival and hazard functions conditioned on u_{iwsz}, respectively, by $S(t \mid u_{iwsz})$ and $\lambda(t \mid u_{iwsz})$. Then

$$f(t \mid X_{n+1} = w, X_n = i, T_n = s, z) = S(t \mid u_{iwsz})\lambda(t \mid u_{iwsz}). \tag{2.1}$$

In particular, putting $u_0 = \{X_{n+1} = i_1, X_n = i_0, T_n = 0, z_0\}$ for some baseline values $i_0 \in \mathcal{S}_1$, $i_1 \in \mathcal{S}$ and z_0, we call $S_0(t) = S(t \mid u_0)$ the baseline survival function. The baseline hazard function $\lambda_0(t)$ is defined by

$$\lambda_0(t) = -\frac{d \log S_0(t)}{dt}.$$

We may set any parametric survival functions as the baseline survival functions, but we will employ generalized Weibull survival function represented by

$$S_0(t) = \exp\biggl(-\int_0^t \lambda_0(x)dx\biggr), \quad \lambda_0(t) = \frac{1}{\theta}\left[1 + \left(\frac{t}{\sigma}\right)^v\right]^{1/\theta - 1} \frac{v}{\sigma}\left(\frac{t}{\sigma}\right)^{v-1}$$

in the application section below.

2.2. The accelerated failure time model

We employ an accelerated failure time modeling technique. Letting $T^*_{n+1} - T^*_n$ be the duration time to the next event under the baseline u_0, we assume that the duration time to the next event at u_{iwsz} changes to $K(u_{iwsz})^{-1}(T^*_{n+1} - T^*_n)$, namely

$$T_{n+1} - T_n = K(u_{iwsz})^{-1}(T^*_{n+1} - T^*_n),$$

where $K(u_{iwsz})$ is an unknown constant that depends on u_{iwsz}.

Then we have

$$\begin{aligned} S(t \mid u_{iwsz}) &= S_0\Big(K(u_{iwsz})t\Big) && (2.2)\\ \lambda(t|u_{iwsz}) &= \lambda_0\Big(K(u_{iwsz})t\Big)K(u_{iwsz}). && (2.3)\end{aligned}$$

We introduce linear functions of u_{iwsz} for $\log K(u_{iwsz})$. Examples of those functions are given as follows.

Example PH1) $\log K(u_{iwsz}) = \beta_0 + \beta_1 i + \beta_2 w + \beta_3^T z + \beta_4 g(s)$, where β_3 is a vector of unknown parameters that has the same dimension as the covariate z, and $g(s)$ is a given function of $s \in \mathbf{R}^1$.

Example PH2) If the transition from state i is possible only to state $i-1$ or $i+1$, we may consider

$$\log K(u_{iwsz}) = \beta_0 + \beta_1 i + \beta_2 x_2 + \beta_3^T z + \beta_4 g(s),$$

where $x_2 = 1$ if $w = i+1$; 0 if $w = i-1$ for $i \in \mathcal{S}_1$.

The survival model developed above is called the accelerated failure time model. It is well known that if Weibull survival function, a special case of the generalized Weibull survival function when $\theta = 1$, is used and $\log K(u_{iwsz})$ is a linear function of i, w, s and z, the accelerated failure time model is equivalent to the Cox proportional hazard model. Note that i, w and s are included in covariates in addition to z in the above development.

2.3. Transition probability

Next we consider transition probabilities. Let $\mathcal{S}_i^* = \{j|P(X_{n+1} = j|X_n = i, T_n = s, z) \neq 0,\ i \in \mathcal{S}_1\}$ and i_0 be the smallest element in $\mathcal{S}_i^*$. Putting

$$g_j(i, s, z) = \log \frac{P\{X_{n+1} = j \mid X_n = i, T_n = s, z\}}{P\{X_{n+1} = i_0 \mid X_n = i, T_n = s, z\}},$$

we consider parametric linear functions of $X_n = i$, $T_n = s$ and z for $g_j(i, s, z)$. Corresponding to the situations in Example PH1 and PH2, examples of such functions are given as follows.

Example TR1) $g_j(i, s, z) = \alpha_{0j} + \alpha_1 i + \alpha_2^T z + \alpha_3 g^*(s)$ for $j \neq i_0$, where $\alpha' s$ are unknown parameters and $g^*(s)$ is a given function of s.

Example TR2) $g_j(i, s, z) = \alpha_0 + \alpha_1 i + \alpha_2 x_2 + \alpha_3^T z + \alpha_4 g^*(s)$, where $x_2 = 1$ if $j = i+1$; 0 if $j = i-1$ for $i \in \mathcal{S}_1$, $\alpha' s$ are unknown parameters and $g^*(s)$ is a given function of s.

It follows that

$$P\{X_{n+1} = j \mid X_n = i, T_n = s, z\} = \frac{\exp\Big(g_j(i,s,z)\Big)}{\sum_{k\in\mathcal{S}_i^*} \exp\Big(g_k(i,s,z)\Big)}. \tag{2.4}$$

Note that $g_{i_0}(i,s,z) = 0$ for any i, s and z. We may assess the effect of $X_n = i$, $T_n = s$ and z on transition probabilities by estimating parameters $\alpha's$ from data based on models given in those examples.

2.4. Likelihood function

Let the process $(X,T) = \{(X_n, T_n); n = 0,1,2,\dots\}$ of an individual with covariate vector z be observed at r time points, giving data (w_i, t_i), $i = 1,2,\dots,r$, during time period $[0,C]$, where C is the censoring time for each individual and r may depend on individual. Then, from assumptions (A) and (B) the log likelihood function of the data from the individual is given as follows.

$$\begin{aligned}\log L = \sum_{n=1}^{r}\Big(&\log f\{t_n - t_{n-1}|X_n = w_n, X_{n-1} = w_{n-1}, T_{n-1} = t_{n-1}, z\}\\ &+ \log P\{X_n = w_n|X_{n-1} = w_{n-1}, T_{n-1} = t_{n-1}, z\}\Big)\\ &+ \log P\{T_{r+1} > C|X_r = w_r, T_r = t_r, z\},\end{aligned}$$

where $f(t|u)$ is given in (2.1) with $S(t|u)$ and $\lambda(t|u)$ given in (2.2) and (2.3),

$$\begin{aligned}&P\{T_{r+1} > C|X_r = a, T_r = s, z\}\\ =& \sum_{k\in\mathcal{S}_a^*} P\{T_{r+1} > C|X_{r+1} = k, X_r = a, T_r = s, z\}P\{X_{r+1} = k|X_r = a, T_r = s, z\}\\ =& \sum_{k\in\mathcal{S}_a^*} \exp\Big(-\int_0^{K(u_{aksz})(C-s)} \lambda_0(u)du\Big)P\{X_{r+1} = k|X_r = a, T_r = s, z\},\end{aligned}$$

and $P\{X_n = w_n|X_{n-1} = w_{n-1}, T_{n-1} = t_{n-1}, z\}$ is given in (2.4). L is the conditional likelihood function given $X_0 = w_0$ and $T_0 = t_0$. Since we are interested in the duration times and transition probabilities, parameters related to initial probabilities are nuisance parameters and they are eliminated by conditioning.

The parameters involved in the duration times and transition probabilities may be estimated by maximizing the log likelihood function for all individuals by using the Newton-Raphson optimization method, more precisely, by using function NLPNRA in SAS/IML Software (SAS Institute).

3. Application

3.1. Background

Government of Japan provides different levels of service to care need individuals based on the severity of their health conditions. The severity is categorized into “need

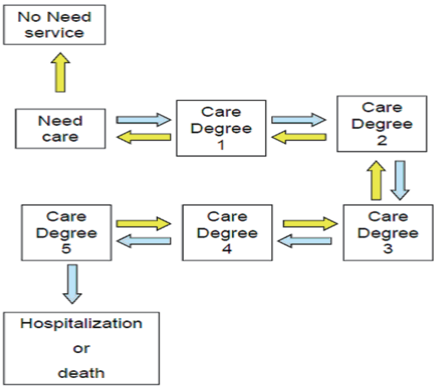

Figure 1: The transition of states

service", "care need degrees one" ,..., and "care need degrees five". Care need individual must take health examination before the service and has to be recognized officially in which category he/she belongs to. The recognition is revised upon the application by each individual. Individuals who are recognized "no need service" and "hospitalization/death" leave from the system. Therefore, the process consists of six transition states, denoted by $\mathcal{S}_1 = \{1, 2, 3, 4, 5, 6\}$ and two absorbing states, denoted by $\mathcal{S}_0 = \{0, 7\}$.

The care service is provided to individuals stayed at nursing homes and stayed at their private houses as well. We focus in this application on people having service at private houses. There were only several data that had two-step transition and we treated them in this application as one-step transition. Thus the transition is possible in this application only to the nearest states. The care system is illustrated in Figure 1. The transition from $i \in \mathcal{S}_1$ to state $i + 1$ is called the *deterioration*, and to state $i - 1$ the *improvement*. The covariates considered are sex, age, family and disease. It was suspected at the beginning of this study that some covariates, such as age, could be related to the transition probability only; and that the disease could be the major influential factor of the duration times to the next event.

3.2. Data

Data were taken from 150 individuals who received care service at private houses for 5 years from April 1, 2000 to March 31, 2005 in a local city of western Japan. Examples of data are as follows. ID1: male, 84 years old, live with family, dementia, the first visit (4/1/00, care degree 2), the first revision (6/24/00, care degree 3), the second revision (12/16/03, care degree 4), the third revision (12/10/05, care degree 3); and ID2: female, 78 years old, live with family, brain disease, the first visit (8/20/02, care degree 1), the first revision (2/18/03, care degree 2), and the second revision (2/25/05, care degree 1). Covariate vector is z=(sex, age, family, disease), where covariates are categorized as follows; sex(0:male, 1:female), age($0 :\leq 80, 1 :> 80$), family(0:living alone, 1:otherwise) and disease (1:dementia or brain disease, 0:otherwise). Covariates age, family and disease are teated as constants.

3.3. Model

We introduced the following model for $K(u_{iwsz})$ that makes it possible to assess the impact of covariates on the duration time from state i to state $i+1$ and from sate i to state $i-1$, separately.

$$\log K(u_{iwsz}) = \begin{cases} \alpha_{10} + \alpha_{11}^T z + \beta_1 i + \gamma_1 g(s) & \text{when} \quad w = i+1 \\ \alpha_{00} + \alpha_{01}^T z + \beta_0 i + \gamma_0 g(s) & \text{when} \quad w = i-1 \end{cases}$$

For the transition probability we employed the following logistic model, which was a special case of (2.4), since there were only two transitions from state $i \in \mathcal{S}_1$, namely, to $i+1$ or $i-1$.

$$\begin{aligned} &\log \frac{P\{X_{n+1} = i+1 | X_n = i, T_n = s, z\}}{P\{X_{n+1} = i-1 | X_n = i, T_n = s, z\}} \\ &\qquad = \alpha_{20} + \alpha_{21}^T z + \beta_1 i + \beta_{21} u_{n1} + \beta_{22} u_{n2} + \beta_{23} u_{n3} + \gamma_2 g^*(s), \end{aligned}$$

where

$$(u_{n1}, u_{n2}, u_{n3}) = \begin{cases} (0,0,0) & \text{if } X_n = 1 \\ (1,0,0) & \text{if } X_n = 2 \\ (0,1,0) & \text{if } X_n = 3 \\ (0,0,1) & \text{if } X_n = 4,5,6 \end{cases} .$$

Here states 4,5 and 6 were pooled since the numbers of observations in these states were small. Functions $g(s)$ and $g^*(s)$ were selected from family $\{s, \log s, \sqrt{s}\ \}$ by checking the goodness of fit of the models to the data by the AIC (Akaike Information Criterion, see, for example, Konishi and Kitagawa (2008)); it was shown that $g(s) = g^*(s) = \sqrt{s}$ was the best fit in the family, but these terms were not significant in the present study. Those covariates included in $K(u_{iwsz})$ were all kept in the model by the request of medical researchers. Just in case we evaluated estimates of selected variables selected by the AIC, and found that the results were not substantially different. The AIC were applied for selecting covariates involved in transition probabilities. The results of the analysis are summarized in Table 1 and 2.

3.4. Results

Table 1 gives estimates, p-values and 95% confidence intervals of parameters in $K(u_{iwsz})$ and estimated parameters of generalized Weibull distribution. The table shows that family and disease are significant in deterioration (p-value=0.000 and 0.000, respectively) and also in improvement (p-value=0.000 and 0.015), showing that the individual living together with other people, or having dementia or brain disease prolong times to improvement and shorten times to deterioration. Also the table shows that the duration time from a state to deterioration/improvement depends on the state significantly, namely, higher states shorten duration times to improvement and prolong duration to deterioration than lower states. It is not easy to understand this findings and we would like to conduct further research to explore the reason. Table 2 lists the results on transition probabilities. The table shows that the odds ratio of the transition to deterioration with respect to improvement from state 2 is significant (p-value=0.006), showing that the transition to improvement from state 2 is approximately 2.4 times more frequent than the transition to deterioration from the same state. Valuable findings are obtained from

Table 1: Estimates, p-values and 95% confidence intervals (CIs) of parameters in $K(u_{iwsz})$.

Covariates		Estimates	p-values	95% CIs	
family	Deterioration	0.729	0.000	0.357	1.102
	Improvement	-0.716	0.000	-1.063	-0.369
sex	Deterioration	0.310	0.075	-0.031	0.651
	Improvement	0.174	0.459	-0.287	0.636
age	Deterioration	0.063	0.697	-0.254	0.380
	Improvement	0.114	0.449	-0.182	0.411
disease	Deterioration	0.531	0.000	0.292	0.771
	Improvement	-0.377	0.011	-0.666	-0.088
state	Deterioration	-0.266	0.000	-0.403	-0.128
	Improvement	0.618	0.000	0.473	0.763
time($\sqrt{s}$)	Deterioration	0.055	0.155	-0.021	0.131
	Improvement	0.007	0.874	-0.075	0.088
Generalized Weibull parameters		Estimates	Std Errors		
	θ	0.146	0.515		
	σ	53.340	564.053		
	v	1.265	0.175		

Table 2: Transition probability; Covariates selected by the AIC, their estimates , odds ratio, p-values and 95% CIs of the estimate.

Covariates	Estimates	Odds ratio	p-values	95%CIs	
Age	0.68	1.97	0.028	0.08 ,	1.29
State 2	-0.88	0.41	0.006	-1.52 ,	-0.25
Time($\sqrt{s}$)	-0.052	0.95	0.466	-0.19 ,	0.09

tables 1 and 2, namely, it is shown that family and disease are significantly related to duration times to the next event, but not related to transition probabilities; on the other hand age is almost related to transition probabilities (p-value=0.051), but not related to duration times to the next event (p-value=0.641 for deterioration; p-value=0.520 for improvement). Although the semi-Markov property was extended slightly by introducing terms $g(s)$ and $g^*(s)$, the impacts of these terms on transition probabilities and duration times were negligible in the present data.

References

Andersen P.K. and Keiding N. (2002). *Multi-state models for event history analysis*, Statistical Methods in Medical Research, 11, 91–115.

Andersen P.K. and Perme M.P. (2008). *Inference for outcome probabilities in multi-state models*, Lifetime Data Analysis, 14, 405–431.

Foucher Y., Giral M., Soulillou J-P and Daures J-P. (2007). *A semi-Markov model for multistate and interval censored data with multiple terminal events. Application in renal transplantation*, Statistics in Medicine, 26, 5381–5393.

Huang J.Z. and Stone C.J. (1998). *The L_2 Rate of Convergence for Event History Regression with Time-dependent Covariates*, Scandinavian Journal of Statistics, 25, 603–620.

Konishi S. and Kitagawa G. (2008). *Information Criteria and Statistical Modeling*, Springer, NY.

Putter H.,Fiocco M and Geskus R.B. (2007). *Tutorial in Biostatistics: Competing risks and multi-state models.*, Statistics in Medicine, 26, 2389–2430.

Received July 26, 2010
Revised August 26, 2010