Methodology Proposal to Estimate Korean ICT Start-ups' Survival: A Discrete-time Proportional Hazard Model

Kyunghoon Kim*

ABSTRACT

This study is concerned with factors affecting start-ups' survival, and with predicting their survival probability. A new proportional hazard model is proposed for analyzing the survival of information and communications technology (ICT) start-ups. This model overcomes the limitations of existing survival models in their utilization of information about firms' activities, stemming from the fact that those models manipulate observations only in a continuous-time horizon. A discrete-time proportional hazard model is proposed to account exhaustively for firms' activities within the study period. To show the superiority of the proposed discrete-time proportional hazard model over a benchmark continuous-time proportional hazard model, both models are applied to a clinical dataset and their log-likelihood values are compared. The proposed model is found to be better in terms of goodness-of-fit and predictive performance.

Key words: Survival analysis, Discrete-time proportional hazard model, Unobserved heterogeneity, ICT Start-ups

JEL Classifications: C1, C5

※ First received, June 19, 2019; Revision received, July 8, 2019; Accepted, July 13, 2019.

* Ph. D. in Management Engineering, Fellow, Department of ICT Strategy Research Korea Information Society Development Institute, E-mail: khkim@kisdi.re.kr

I. INTRODUCTION

In general, the survival time of start-ups tends to be shorter than usual. In case of US start-ups, the failure rate in the first year is about 40%, and it is about 90% within 10 years from the establishment year (Timmons, 1990). In most countries that belong to the organization for economic co-operation and development (OECD), more than half of the start-ups closed within five years (OECD, 2015). In Mata & Portugal (1994), more than one-fifth of the start-ups in Portugal closed in their establishment year, and only 40% survived for more than seven years. Since the growth of start-ups is directly aligned with the maintenance of national competitiveness, it is very important to investigate factors that affect whether start-ups survive or not and predict their survival probability. It is especially crucial to analyze the survival of start-ups in the information and communications technology (ICT) industry, because not only is the contribution of ICT to gross domestic product (GDP) non-negligible, but the related market is also too competitive.

However, few studies analyzed the survival behavior for ICT start-ups in South Korea. Some studies investigated the relationship between firms' characteristics such as accounting data, intellectual property, chief executive officer's characteristics, and their survival times, while others analyzed whether the government's support policy affects their performances. They did not focus on ICT start-ups but on small and medium enterprises (SMEs) in South Korea (Lee & Shin, 2005; Lim et al., 2008). Moreover, existing survival models do not adequately explain their hazard patterns. For example, models that make use of the conventional Cox proportional hazard model, widely used in survival analysis, are unable to utilize information about a firm's activity sufficiently, because observations in those models are manipulated only in the continuous-time horizon (Shin et al., 2017).

Therefore, this study aims to propose a new proportional hazard model for analyzing ICT start-ups' survival to overcome the limitation that existing survival models are unable to utilize information about a firm's activity sufficiently. To utilize what firms do within the study period as much as possible, we develop the discrete-time proportional hazard model, first proposed by Thompson (1977).

The remainder of this study is organized as follows. In Section II, we review previous survival analysis studies. In Section III, we develop the modeling framework and apply it to the clinical dataset of cancer patients to prove that the proposed model is superior to the benchmark model. In Section IV, we conclude with a summary of this study and its contributions.

II. LITERATURE REVIEW

The most famous model in survival analysis is the proportional hazards regression model, first proposed by Cox (1972). The Cox model has been used widely to specify a linear relationship between the hazard or survival rates and covariates in a variety of fields, such as engineering, economics, and sociology. Many studies, especially in marketing, have adopted the proportional model to characterize the purchase-timing behavior of households (Jain & Vilcassim, 1991; Vilcassim & Jain, 1991). In biometrics, some studies have also used this method to model an individual's lifetime (Prentice & Kalbfleisch, 1979; Lancaster, 1979; Hakulinen & Tenkanen, 1987). The rationale of the Cox model is that a subject's hazard probability—the instantaneous probability of dying—consists of two components: baseline hazard, which represents the subjects' intrinsic temporal survival pattern, and the covariate function, which explains the influence of time-varying covariates such as research and development (R&D) intensity. In the Cox model, the hazard function is multiplicatively decomposed into these two components. An alternative specification of covariate effects is the additive risk model, first proposed by Aalen (1980), in which time-varying covariates are allowed to influence survival time by acting in an additive way.

These types of proportional hazard models, however, overlook the possibility of unobserved heterogeneity. As proved in Lancaster (1992), if unobserved heterogeneity is ignored, the estimates could be spurious. Lancaster (1979) used a Weibull-gamma mixture or a finite mixture model based on Bayesian methods to capture the unobserved heterogeneity. Moreover, many studies proved that an incorrect assumption about the unobserved heterogeneity distribution in survival models could incur severe consequences (Heckman & Singer, 1984; Bretagnolle & Huber-Carol, 1988; Hougaard et al., 1994; Baker & Melino, 2000). On the other hand, Nicoletti & Rondinelli (2010) showed that there are no major biases in estimating the survival and expected duration functions when neglecting or mis-specifying the unobserved heterogeneity distribution. In the context of individual frailty, Vaupel et al. (1979) first proposed a frailty model to deal with the issue of heterogeneity, assuming that more frail individuals have a tendency to survive less than those who are less frail. Duchateau & Janssen (2007) also introduced some frailty models using gamma, Weibull, and log-normal parametric distributions. Jun et al. (2015) suggested a framework for predicting individual survival times using a Weibull-gamma mixture model with individual covariates.

Although the survival time from the first observed time until closure is of primary interest in survival analysis, what firms do (e.g., to increase R&D investment) within the study period also plays an important role in accounting for the hazard function. As an alternative model to the continuous-time proportional hazard model, the discrete-time proportional hazard model, first proposed by Thompson (1977), is used to factor in a firm's activity with the observation that it still exists. A discrete-time model is better than a continuous-time model in terms of model validity, because many studies have made observations only in the discrete-time horizon. In marketing, many researchers, including Gupta (1991), Helsen & Schmittlein (1993), and Seetharaman & Chintagunta (2003), used this discrete-time proportional hazard model to explain shopping trips with non-purchase of a product. Schweidel et al. (2008) adopted the discrete-time proportional hazard model to capture consumers' service retention behavior, whereas Nam et al. (2008) used it to predict firms' bankruptcy.

Consequently, we propose a new survival model, called the discrete-time proportional hazard model, to overcome a limitation based on the Cox proportional hazard model. Specifically, we change the time horizon in the hazard function from continuous to discrete by replacing the integral equation into a summation formulation and use the Weibull-gamma mixture model into the baseline hazard function to capture unobserved heterogeneity as in Jun et al. (2015). More details are presented in Section III.

III. MODEL DEVELOPMENT

Typically, the objective of survival analysis is to examine the relationship between survival time and individual covariates or explanatory variables. Cox (1972) proposed the proportional hazard model, which takes the hazard function for firm i given x_t written as

$$h_i(t,x_t)= h_i(t)\exp\left(x_t^T\delta\right), \tag{1}$$

where $h_i(t)$ is a baseline hazard function that stands for the risk to firm i with $x_t = 0$, and $\exp\left(x_t^T\delta\right)$ is the proportionate increase or reduction in risk with x_t. x_t represents the column vector of time-varying covariates that can affect the length of survival time including accounting variables such as asset, liability, capital, sales, operating cost, and so on. δ is the row vector of the parameters corresponding to x_t.

The Cox proportional hazard model is sometimes called a semi-parametric model because the baseline hazard function is usually assumed to be model-free. However, we suppose that the baseline hazard in my model follows statistical distribution such as an exponential or Weibull distribution, as we mentioned before, to capture the unobserved heterogeneity in individual firms' patterns of hazard. Therefore, we first specify the baseline hazard with parametric distribution and unobserved heterogeneity before developing my proposed model—the discrete-time proportional hazard model.

1. Baseline Hazard

A variety of parametric specifications including exponential, Weibull, Erlang-2, log-logistic, and expo-power distributions have been used for the baseline hazard. Even though the log-logistic or expo-power distributions have been shown to be the most appropriate to specify the baseline hazard (Seetharaman & Chintagunta, 2003), the Weibull distribution is most widely used because of its flexibility and ease of interpretation[1]. Therefore, we use the Weibull distribution as in Jun et al. (2015).

The survival probability and probability density function of the Weibull distribution with two parameters, λ and c, are given by

$$S_i(t) = \exp(-\lambda t^c), \tag{2}$$

$$f_i(t) = \lambda c t^{c-1} \exp(-\lambda t^c). \tag{3}$$

We allow for differences in rate parameter λ across firms by using their covariates and reflecting the unobserved heterogeneity. That is, the rate parameter is assumed to be a multiplicative form of two components, λ_{1i} and λ_{2i}, which represent individual covariate effects and unobserved heterogeneity, respectively.

The first component, λ_{1i}, is represented as a linear combination of covariates as follows:

$$\lambda_{1i} = \exp(z_i^T \gamma), \tag{4}$$

where z_i is a vector containing covariates related to the time-invariant characteristics of firm i, and γ is a vector of all parameters associated with the covariates. The time-invariant covariates comprise the business sector, registry, target marketplace, and founder's socio-demographic information such as sex, university major, and education level. Since λ_{2i} plays a role in the rate parameter as an intercept, the constant term is excluded from z_i to avoid an identification problem.

When firms are assumed to be taken from a heterogeneous population, unobserved heterogeneity should be dealt with in the baseline hazard. If heterogeneity is ignored, the estimates could be spurious (Lancaster, 1992). Therefore, we set the second component λ_{2i} to capture the unobserved heterogeneity in the likelihood of survival. As in Lancaster (1979), we assume that λ_{2i} is drawn from a gamma distribution.

[1] Since the exponential distribution has a single parameter, it only explains a constant hazard rate. On the other hand, the Weibull distribution can account for both the increasing and decreasing hazard rates by manipulating two parameters.

$$g(\lambda_{2i}|\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda_{2i}^{\alpha-1}\exp(\beta\lambda_{2i}), \tag{5}$$

where α and β are the shape and rate parameters of the gamma distribution, respectively. We use the gamma distribution not only because it is the conjugate prior for the Weibull distribution but also for its flexibility. By integrating the probability density function of the Weibull distribution(Equation (3)) over Equation (5), the probability density function and survival probability for firm i, given c, γ, α, and β, can be derived as follows:

$$f_i(t|c,\gamma,\alpha,\beta)=\int_{\lambda_{2i}} f_i(t|c,\gamma,\lambda_{2i})g(\lambda_{2i}|\alpha,\beta)d\lambda_{2i} \tag{6}$$

$$=ct^{c-1}\exp\left(z_i^T\gamma\right)\left[\frac{\alpha}{\beta+t^c\exp\left(z_i^T\gamma\right)}\right]\left[\frac{\beta}{\beta+t^c\exp\left(z_i^T\gamma\right)}\right]^{\alpha},$$

$$S_i(t|c,\gamma,\alpha,\beta)=\left[\frac{\beta}{\beta+t^c\exp\left(z_i^T\gamma\right)}\right]^{\alpha}. \tag{7}$$

Therefore, the baseline hazard can be specified as follows:

$$h_i(t|c,\gamma,\alpha,\beta)=\frac{f_i(t|c,\gamma,\alpha,\beta)}{S_i(t|c,\gamma,\alpha,\beta)}=ct^{c-1}\exp\left(z_i^T\gamma\right)\left[\frac{\alpha}{\beta+t^c\exp\left(z_i^T\gamma\right)}\right]. \tag{8}$$

2. Continuous-time Proportional Hazard Model (Benchmark Model)

In the case of the continuous-time proportional hazard model, the hazard function can be expressed by

$$h_i(t,x_t)=\frac{f_i(t,x_t)}{S_i(t,x_t)}, \tag{9}$$

where $f_i(t,x_t)$ is the probability density function, and $S_i(t,x_t)$ is the survival function, defined as one minus the cumulative distribution function (i.e.,

$1-F_i(t,x_t))$ for firm i at time t. Then, Equation (9) is equivalent to

$$h_i(t)\exp\left(x_t^T\delta\right)=\frac{f_i(t,x_t)}{S_i(t,x_t)}\ . \tag{10}$$

The fact that $f_i(t,x_t)=\frac{d}{dt}F_i(t,x_t)=-\frac{d}{dt}S_i(t,x_t)$ yields

$$\int h_i(t)\exp\left(x_t^T\delta\right)dt=\int\frac{f_i(t,x_t)}{S_i(t,x_t)}dt=-\log S_i(t,x_t)\ . \tag{11}$$

Therefore, the survival and probability density functions of the continuous-time proportional hazard model for firm i at time t can be written as follows:

$$S_i(t,x_t)=\exp\left[-\int_0^t h_i(s)\exp\left(-x_s^T\delta\right)ds\right]\ , \tag{12}$$

$$f_i(t,x_t)=h_i(t)\exp\left(-x_t^T\delta\right)\exp\left[-\int_0^t h_i(s)\exp\left(-x_s^T\delta\right)ds\right]\ . \tag{13}$$

The parameters of the continuous-time proportional hazard model are estimated using a maximum likelihood estimation. If a firm has closed before the cutoff date (i.e., the observation is uncensored), its contribution to the likelihood function is the probability density function of the survival time. On the other hand, for a firm that still exists at the cutoff date (i.e., the observation is censored), all the researchers know is that the survival time exceeds the difference between its starting date and the cutoff date. Therefore, its contribution to the likelihood is considered to be the survival probability for the number of months or years before the censoring occurred. Consequently, the log-likelihood function can be represented as follows:

$$LL=\sum_i^n\delta_i\log f_i(t)+\sum_i^n(1-\delta_i)\log S_i(t)\ . \tag{14}$$

3. Discrete-time Proportional Hazard Model (Proposed Model)

On the other hand, in the discrete-time proportional hazard model, the integral of the hazard function in the continuous-time model (i.e., Equation(12)) is replaced by the summation. That is, the survival function can be written as follows:

$$S_i^{(d)}(t,x_t)=\exp\left[-\sum_{s=1}^{t}\exp\left(x_s^T\delta\right)\int_{s-1}^{s}h_i(u)du\right], \tag{15}$$

where s is the discrete time measured in months[2]. This formulation is more plausible for two reasons. First, the time-varying covariates affect a firm's survival probability at a discrete time and usually vary from month to month, not within a given month. Second, a firm's status varies every month even though it still exists. That is, every month, the firm survives under uncertainty with the probability of closure determined by the covariates.

Then, the discrete-time hazard function, defined as the probability of closure for firm i in month t since the previous month $t-1$, given x_t, can be written as follows:

$$h_i^{(d)}(t,x_t)=\frac{P_i^{(d)}(t,x_t)}{S_i^{(d)}(t-1,x_{t-1})}=1-\frac{S_i^{(d)}(t,x_t)}{S_i^{(d)}(t-1,x_{t-1})}, \tag{16}$$

where $P_i^{(d)}(t,x_t)$ is the probability mass function, defined as the difference between the survival function in month t and $t-1$[3]. Substituting Equations (8) and (15) into Equation (14) yields the following discrete-time proportional hazard model:

$$\begin{aligned} h_i^{(d)}(t,x_t)&=1-\exp\left[-\exp\left(x_i^T\delta\right)\int_{s-1}^{s}h_i(u)du\right] \\ &=1-\exp\left\{-\frac{\exp\left(x_i^T\delta\right)}{\exp\left(z_i^T\gamma\right)}\alpha\log\left[\frac{\beta+t^c\exp\left(z_i^T\gamma\right)}{\beta+(t-1)^c\exp\left(z_i^T\gamma\right)}\right]\right\}. \end{aligned} \tag{17}$$

[2] To distinguish this from the continuous-time proportional hazard model, hereafter, a superscript (d) is added to the discrete-time model.

[3] $P_i^{(d)}(t,x_t)=S_i^{(d)}(t-1,x_{t-1})-S_i^{(d)}(t,x_t)$

The parameters in Equation (17) can be estimated using a maximum likelihood estimation. Different from the continuous-time proportional hazard model, this estimation does not need to deal with the censoring issue because we consider the firm's status (or covariates) every month, regardless of whether the firm is closed. Therefore, the log-likelihood function can be represented as follows:

$$LL = \sum_i \left[\sum_{t=1}^{T_i} \phi_t^i \log\left(h_i^{(d)}(t, x_t)\right) + \sum_{t=1}^{T_i} \left(1 - \phi_t^i\right) \log\left(1 - h_i^{(d)}(t, x_t)\right) \right] , \qquad (18)$$

where ϕ_t^i represents the indicator variable, which is set to one if firm i closed in month t and zero otherwise, and T_i is the total spell from the first observed month for firm i.

4. Empirical Results

To prove that the proposed model—the discrete-time proportional hazard model with unobserved heterogeneity—is superior to the benchmark model—the continuous-time proportional hazard model with unobserved heterogeneity—we apply the proposed model to clinical survival data retrieved from the Surveillance, Epidemiology, and End Results Program database[4] of the National Cancer Institute. Specifically, the data consist of information from nine registries in the United States about the survival times of patients diagnosed with respiratory cancer from 1988 to 2012 , which is approximately 10% of the entire US population. The data are right-censored, with the follow-up cutoff date fixed at December 31, 2012. The total number of patients is 189,938. Among them, 186,089 patients were diagnosed once; 111,929 patients are known to have died before the cutoff date (i.e., the observations are uncensored), and the fate of 74,160 patients after the cutoff date is unknown (i.e., the observations are censored). Similarly, 3,849 patients were diagnosed two or more times; 1,674 patients are uncensored, whereas 2,175 patients are censored.

There are two reasons why we use medical history of respiratory cancer patients instead of Korean ICT start-ups. First, there is no appropriate data to analyze the

[4] http://www.seer.cancer.gov/. Released April 2014, based on the November 2013 submission.

survival behavior of Korean ICT start-ups. Even though "ICT venture panel data" constructed by the Korea Information Society Development Institute covers survival times for Korean ICT venture companies, it is susceptible to sampling bias (Jo et al., 2018). Specifically, as the samples are selected from venture companies in ICT industry, established between the year 2012 and 2015 and still alive in 2016, the survival time must be longer than expected. For example, all companies established in 2012 survives more than three years. As shown in <Table 1>, among 1,085 samples, only 10 companies (0.9%) closed within three years, and 70 companies (6.5%) closed during the study period. In addition, the average survival time until closure is 4.2 years. These statistics are not consistent with those from the existing studies such as Timmons (1990) or OECD (2015).

<Table 1> Survival time distribution (ICT venture panel data)

Survival Time	Number of Companies	Descriptive Statistics (Survival Time)		
		Mean	Max	Min
< 1 year	0 (0.0%)	4.20 years	6.08 years	1.57 years
< 2 years	2 (0.2%)			
< 3 years	10 (0.9%)			
< 4 years	26 (2.4%)			
< 5 years	54 (5.0%)			
< 6 years	68 (6.3%)			
< 7 years	70 (6.5%)			

Note: Total number of companies is 1,085. Numbers in parentheses indicate the ratio of number of closed companies to total number of companies in the data.

Source: Jo et al. (2018)

Second, respiratory cancer patients are similar to start-ups in terms of survival distribution. As aforementioned, the failure rate of US start-ups in the first year is about 40%, and it is about 90% within 10 years. In addition, half of the start-ups from the OECD countries closed within five years. Similarly, only 15.2% of respiratory cancer patients in Korea survived within 10 years (Ministry of Health & Welfare, 2018). Moreover, the discrete-time approach is valuable in the clinical data because patients receive medical treatment such as surgery or radiation several times within the study period. This information contributes to explaining their hazard function more accurately if they are used in the discrete-time proportional hazard model.

Time-invariant covariates z_i in the baseline hazard comprise individual sex, race, registry, and cohort, and time-varying covariates x_t that can affect the length of survival time include age, tumor stage, and medical treatments, as shown in <Table 2>.

<Table 2> Description of time-invariant and time-varying covariates

Group	Covariates	Description
Time-invariant Covariates	sex	1, if the subject is male 0, otherwise
	$race^j$	1, if the subject is of race *j*, where *j*=1 (white), or 2 (black) 0, otherwise
	reg^j	1, if the subject is registered in registry *j*, where *j*=1 (San Francisco-Oakland), 2 (Connecticut), 3 (Detroit), 4 (Hawaii), 5 (Iowa), or 6 (New Mexico) 0, otherwise
	$cohort$	subject's diagnosis year - age
Time-varying Covariates	age_t	subject's age at time *t*
	$stage_t$	1, if the subject's stage at time *t* is in-situ 2, if the subject's stage at time *t* is local/regional 3, if the subject's stage at time *t* is distant
	$surg_t$	1, if the subject has received surgery at time *t* 0, otherwise
	rad_t	1, if the subject has received radiation at time *t* 0, otherwise

<Table 3> reports parameter estimates for the continuous and discrete-time proportional hazard models. Most of the coefficients are significant at the 1% level with the expected signs. Since the time-invariant covariates and hazard rate have a positive relationship, in the continuous-time proportional hazard model, it can be understood that (i) male subjects have higher risk than female subjects do, (ii) other races have higher risk than white or black subjects do, and (iii) subjects in more recent cohorts have lower risk than those in previous cohorts do. On the other hand, in the discrete-time proportional hazard model, since

the time-invariant covariates and hazard rate have a negative relationship, the estimation results are consistent with those in the continuous-time proportional hazard model, except for the race covariates and some registry covariates. For the time-varying covariates, the signs of parameters are found to be the same in both models. (i) For the stage covariate, the hazard rate is higher for subjects at a distant stage than in situ or at the local stage. (ii) In the case of medical treatments, subjects who have undergone surgery or radiation have lower risk than those who have not. (iii) Lastly, the fact that the age covariate estimate is negative shows that older subjects have lower risk than younger subjects do. This seems unusual, but the value is too small to conclude that this negative effect overwhelms other effects.

<Table 3> Estimation results

Parameter			Continuous-time Proportional Hazard Model		Discrete-time Proportional Hazard Model	
			Estimate	**(*t*-statistics)**	**Estimate**	**(*t*-statistics)**
Shape parameter			1.18***	(18.92)	1.62***	(221.07)
Time-invariant Covariates	Sex		0.10***	(7.25)	-0.27***	(-29.30)
	Race	White	-0.91***	(-6.00)	-0.21***	(-10.65)
		Black	-0.90***	(-5.64)	-0.35***	(-14.82)
	Registry	San Francisco-Oakland	-0.67***	(-7.89)	0.06**	(2.35)
		Connecticut	-0.64***	(-7.92)	0.14***	(5.90)
		Detroit	-0.60***	(-8.09)	0.12***	(5.22)
		Hawaii	-1.25***	(-7.09)	-0.06*	(-1.88)
		Iowa	-0.30***	(-5.90)	0.16***	(5.34)
		New Mexico	-0.49***	(-6.47)	-0.13***	(-4.59)
	Cohort		-0.05***	(-13.80)	0.02***	(39.18)
Time-varying Covariates	Age		-0.07***	(-9.72)	-0.01***	(-15.82)
	Stage		0.55***	(19.86)	1.35***	(207.63)
	Medical Treatment	Surgery	-2.34***	(-16.12)	-3.87***	(-13.68)
		Radiation	-0.14***	(-3.81)	-5.19***	(-19.55)
Unobserved Heterogeneity	Shape parameter		0.96***	(4.22)	0.00	(0.10)
	Rate parameter		1.04	(1.16)	3.33***	(55.56)
Log-likelihood(*n*=189,938)			-495,985		-475,119	

Note: *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

The shape parameter estimates in both models are significantly greater than one, which means that subjects exhibit positive duration dependence (or increasing hazard rate) in their probabilities of dying. That is, subjects are more likely to die the longer they are alive. If the unobserved heterogeneity is ignored in the baseline hazard, as in Follmann & Goldberg (1988), a spurious decreasing hazard rate is obtained. Since the log-likelihood value is higher in the discrete-time proportional hazard model, we conclude that this model explains subjects' hazard patterns better than the continuous-time proportional hazard model.

To compare the predictive performance of the discrete and continuous-time proportional hazard models, we divide all subjects into two sub-samples—hold-in and hold-out—and assess the log-likelihoods of each model. Specifically, we estimate the parameters of the two models using 80% (151,950/189,938) of all subjects and then compute the log-likelihoods using the remaining 20% (37,988/189,938). <Table 4> shows that the log-likelihood of the discrete-time proportional hazard model is higher in both the hold-in and hold-out samples than that of the continuous-time proportional hazard model.

<Table 4> Predictive performance: log-likelihood value

Group	Continuous-time Proportional Hazard Model	Discrete-time Proportional Hazard Model
Hold-in sample (*n*=151,950)	-396,799	-371,989
Hold-out sample (*n*=37,988)	-99,361	-95,661

IV. CONCLUDING REMARKS

In this study, we proposed a new proportional hazard model for analyzing ICT start-ups' survival to overcome the limitation that existing survival models are unable to utilize information about a firm's activity sufficiently. This is because observations in those models are manipulated only in the continuous-time horizon. To utilize what firms do within the study period as much as possible, we developed the discrete-time proportional hazard model, and we suggested the Weibull-gamma mixture model to capture the unobserved heterogeneity between start-ups. To verify whether my proposed model is better than an existing survival model , we applied two models—the discrete-time proportional hazard model as the proposed model and the continuous-time proportional hazard model as the benchmark model—to clinical data and compared their log-likelihood values. As a result, in both the hold-in and hold-out samples, the derived log-likelihood values that measure goodness-of-fit and predictive performance in the proposed model were higher than those in the benchmark model.

Despite the demonstrated performance of the proposed model, this study has two limitations. First, we did not use Korean ICT start-ups' data as an empirical study. This is because there is no appropriate data to analyze the survival behavior of Korean ICT start-ups. Instead, we used medical history of respiratory cancer patients, as it is similar to start-ups in terms of survival distribution. The discrete-time approach can be valuable for studying clinical data, as patients receive medical treatment such as surgery or radiation several times within the study period. Therefore, it seems reasonable to use medical history as a proxy of ICT start-ups. Second, this study does not guarantee the same empirical result when using other data. In other words, the fact that the results are derived as expected can be interpreted to mean that the proposed model is well fitted for the clinical data. For example, the assumption of Weibull-gamma mixture in modeling the baseline hazard can be adequate only for our dataset. As many studies have mentioned that an incorrect assumption about the unobserved heterogeneity distribution in survival models can incur severe consequences (Heckman & Singer, 1984; Bretagnolle & Huber-Carol, 1988; Hougaard et al., 1994; Baker & Melino, 2000), the empirical results depend on harmony between the data and the assumption of unobserved heterogeneity distribution. That is, one needs to be more careful in modeling unobserved heterogeneity in future research.

Nevertheless, we believe that the proposed model derived from this study can shed light on the study of ICT start-ups' survival in South Korea. As aforementioned, it is very important to investigate factors that affect whether start-ups survive or not, because the growth of a start-up is directly aligned with the maintenance of the country's competitiveness. This is especially crucial in the ICT industry, because not only is the contribution of ICT to GDP is non-negligible but the related market is also too competitive. However, there are few studies on this topic, and data for analyzing the survival behavior of ICT start-ups in South Korea and the existing survival models do not show good predictive performance. Since the proposed model is established for analyzing survival or hazard patterns more adequately than existing models, it is also expected to contribute to survival studies on South Korea's ICT industry if there exist sufficient data on ICT start-ups' survival.

REFERENCES

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Mathematical Statistics and Probability Theory, Lecture Notes in Statistics*, *Vol. 2*, Springer-Verlag, New York, 1-25.

Baker, M., & Melino, A. (2000). Duration depencence and nonparametric heterogeneity: a Monte Carlo study. *Journal of Econometrics*, *96*, 357-393.

Bretagnolle, J., & Huber-Carol, C. (1988). Effect of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, *15*, 125-138.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, *34(2)*, 187-220.

Duchateau, L., & Janssen, P. (2007). *The frailty model.* Springer.

Follmann, D. A., & Goldberg, M. S. (1988). Distinguishing heterogeneity from decreasing hazard rates. *Technometrics*, *30(4)*, 389-396.

Gupta, S. (1991). Stochastic models of interpurchase time with time-dependent covariates. *Journal of Marketing Research*, *28*, 1-15.

Hakulinen, T., & Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *36(3)*, 309-317.

Heckman, J. J., & Singer, B. (1984). Econometric duration analysis. *Journal of Econometrics*, *24*, 63-132.

Helsen, K., & Schmittlein, D. C. (1993). Analyzing duration times in marketing: evidence for the effectiveness of hazard rate models. *Marketing Science*, *11*, 395-414.

Hougaard, P., Myglegaard, P., & Borch-Johnsen, K. (1994). Heterogeneity models of disease susceptibility with an application to diabetic nephropathy, *Biometrics*, *50*, 1178-1188.

Jain, D. C., & Vilcassim, N. J. (1991). Investing household purchase timing decisions: a conditional hazard function approach. *Marketing Science*, *10*, 1-23.

Jo, Y., Kim, K., Lee, E., Lee, D., & Choi, C. (2018). Analysis of causes of ICT venture startups' success and failure and ways to strengthen the competitiveness of the venture startup ecosystem: via the ICT venture panel. *Korea Information Society Development Institute*, *18-07-03.*

Jun, D. B., Kim, K., Park, M. H. (2017). Forecasting annual lung and bronchus cancer deaths using individual survival times. *International Journal of Forecasting*, *32(1)*, 168-179.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, *47(4)*, 939-956.

____________ (1992). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

Lee, B. K., & Shin, K. C. (2005). The determinants of new firms survival: an empirical analysis using hazard model. *Kukje Kyungje Yongu*, *11(1)*, 131-154.

Lim, C., Lee, Y., Lee, K., Kim, J., Bae, Y., & Kim, S. (2008). An analysis for Korean venture survival. *Science and Technology Policy Institute*, *2008-11.*

Mata, J., & Portugal, P. (1994). Life duration of new firms. *Journal of Industrial Economics*, *42(3)*, 227-245.

Ministry of Health, & Welfare (2018). 2016 National Cancer Registry Statistics.

Nam, C. W., Kim, T. S., Park, N. J., & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, *27*, 493-506.

Nicoletti, C., & Rondinelli, C. (2010). The (mis)specification of discrete duration models with unobserved heterogeneity: A Monte Carlo study. *Journal of Econometrics*, *159(1)*, 1-13.

OECD (2015). *Entrepreneurship at a Glance 2015*. OECD Publishing, Paris.

Prentice, R. L., & Kalbfleisch, J. D. (1979). Hazard rate models with covariates. *Biometrics*, *35(1)*, 25-39.

Schweidel, D. A., Fader, P. S., & Bradlow, E. T. (2008). Understanding service retention within and across cohorts using limited information. *Journal of Marketing*, *72*, 82-94.

Seetharaman, P. B., & Chintagunta, P. K. (2003). The proportional hazard model for purchase timing: a comparison of alternative specifications. *Journal of Business & Economic Statistics*, *21(3)*, 368-382.

Shin, K., Park, G., Choi, J. Y., & Choy, M. (2017). Factors affecting the survival of SMEs: A study of biotechnology firms in South Korea. *Sustainability*, *9(1)*, 108.

Thompson, W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics*, *10*, 411-431.

Timmons, J. A. (1990). *New venture creation: entrepreneurship in the 1990s*. Homewood, IL, Irwin.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*, 439-454.

Vilcassim, N, J., & Jain, D. C. (1991). Modeling purchase timing and brand switching behavior incorporating explanatory variables and unobserved heterogeneity. *Journal of Marketing Research*, *28*, 29-41.