International Journal of Education and Human Developments, Vol. 9 No. 1; March 2023
ISSN 2415-1270 (Online), ISSN 2415-1424 (Print)
Published by Center for Global Research Development

Research Tools for Analysis of Dichotomous Outcomes in Education: A Comparison of Discrete (Logit) and Continuous (Clog-Log) Survival Analysis

Sunha Kim, PhD
Associate Professor
Department of Counseling, School, and Educational Psychology
Department of Learning and Instruction
Graduate School of Education
SUNY at Buffalo
423 Baldy Hall, Buffalo, NY 14260
USA

Mido Chang, PhD
Professor
Department of Counseling, Recreation, & School Psychology
College of Arts, Sciences & Education
Florida International University
11200 SW 8th St. ZEB 250B, Miami, FL 33199
USA

Abstract

This study aimed to provide a guideline for the proper use of a survival model in an educational research field in which discrete and continuous models often need clarification. To achieve the goal, we compared discrete- (logit) and continuous-time (complementary log-log) survival models by considering the hazard rate. We simulated data for various combinations of time metrics, censoring proportions, and sample size. The study results recommend discrete models for cases with large time metrics, small proportions of censored observations, and small samples. The study highlights the importance of adopting a proper model in using survival analyses.

Keywords: survival analysis; discrete (logit) model, continuous (clog-log) model, censoring proportion, hazard rate, sample size, time metric

1. Introduction

Survival analysis is an advanced statistical method that deals with dichotomous outcomes in longitudinal data. Survival analysis has been prevalently adopted by medicine and biology to estimate the time to events such as death, recovery from disease, or treatment responses. (Klein & Moeschberger, 2003). Recently, survival analysis has been used in education fields by modeling the occurrence of events, such as student dropouts or teacher attrition, in a longitudinal time frame (Kelly, 2004; Kirby, Berends & Naftel, 1999; Ma & Willms, 1999; Murphy, Gaughan, Hume & Moore, 2010; Plank, DeLuca, & Estacion, 2008). Survival analysis estimates a hazard function, also called a conditional risk, such that a target event will occur, given that the target event has not happened yet. It also uses two-event estimation methods: discrete-time and continuous-time estimations (Singer & Willet, 2003). The discrete-time estimation uses the time to a target event in a large time metric, such as a year, quarter, or month. Continuous time estimation records an event time with a precise and fine metric, such as an hour, day, or week.

Despite the clear distinction between discrete- and continuous-time estimation, literature reviews show that many studies have adopted either of them without clear distinctions (Calcagno, Crosta, Bailey & Jenkins, 2007; Doyle, 2006; Donaldson & Johnson, 2010; Kahn & Schwalbe, 2010; Kelly, 2004; Kirby, Berends & Naftel, 1999; Ma & Willms, 1999; Murphy, Gaughan, Hume & Moore, 2010; Murtaugh, Burns & Schuster, 1999; Singer, 1992; Singer, Davidson, Graham & Davidson, 1998). Adopting a survival model without an empirical and theoretical understanding can be problematic because an incorrect model may result in biased estimates and incorrect conclusions.

In response, this study attempts to provide analysis results using various simulated data and attempts to suggest a guideline for selecting a discrete or continuous model for survival analysis. Specifically, this study compared logit and complementary log-log (clog-log) models, representing discrete and continuous models. This study paid particular attention to a hazard rate, which has been identified as an essential variable in distinguishing between the two models (Hess, 2009; Hosmer & Lameshow, 1999; Singer & Willet, 2003). Furthermore, the study included three factors closely associated with hazard rate. These factors are time metrics, censoring proportion, and sample size. By combining the three factors, this study simulated a series of data to employ both logit and clog-log models. The parameter estimates and fit statistics resulting from the two model analyses were compared to investigate the discrepancy between the two estimation methods.

To achieve the goal, the study adopted the following overarching research question:

- How different are the outcomes of the logit and complementary log-log survival models in various data conditions?

The study also addressed the next three research questions:

- Do hazard rates lead to discrepancies between the two model outcomes? If they do, in which condition of hazard rates does a discrepancy appear?
- Do time metrics influence discrepancies between the two model outcomes? If they do, in which condition of time metrics does a discrepancy show?
- Do censoring proportions associate with discrepancies between the two model outcomes? If they do, in which condition of censoring proportions does a discrepancy appear?
- Do sample sizes relate to discrepancies between the two model outcomes? If they do, in which condition of sample sizes does a discrepancy reveal?

2. Literature review

2.1. Survival analysis

Using longitudinal data, survival analysis estimates whether, when, and why an event of interest (target event) occurs (Singer & Willet, 2003). When conducting a survival analysis, researchers should consider not only the onset of the target event but also the time metrics, tied observations, and censored observations. The time metrics can be measured in either discrete or continuous time units. Discrete-time units record the time passage in a large (broader) metric (i.e., semester or year), while continuous time units record the time passage in a fine (more precise) metric (i.e., day, hour, or minute). The time metrics also determine the tied observations, which are defined as two or more observations that occur at the same time. Therefore, when there are smaller time units, there will be fewer tied observations. Censored observations refer to cases in which the target event does not occur during the study's data collection. It is important to note that survival analysis takes censored observations into account in its modeling, which makes its design stronger than other longitudinal analyses.

2.2. Discrete-time vs. continuous-time hazard models

Discrete-time survival analysis estimates the risk (probability) of the target event's occurrence in comparatively larger time units. The risk is estimated as a conditional probability that the event of interest will occur. The discrete-time survival analysis assumes that two or more observations will occur simultaneously, as it uses broad time metrics. Thus, the discrete-time survival analysis is recommended in the case of strong ties, which does not lead to biased estimates for those conditions.

Continuous time estimation records the occurrence of events in fine units such as minutes, hours, or days. Among continuous-time survival models, the most popular model is the Cox proportional hazard model (Singer & Willet, 2003). According to the Cox proportional hazards model, a hazard is estimated as an instantaneous change in the occurrence rate of the event. When building a Cox proportional hazard model, one should be very careful in dealing with ties because the Cox model is very sensitive to ties and may lead to invalid results. Common methods to manage ties in continuous-time survival analysis include the exact method, the Breslow-Peto approximation, and the Efron approximation (Hertz-Picciotto & Rockhill, 1997). The exact method treats all the possible combinations of observations in ranking tied observations. The Breslow-Peto approximation randomly posits a sequential occurrence of tied observations. The Efron approximation assumes all the possible rankings of tied observations but adopts simple computations. Among these three methods, the exact method is the most recommended, followed by the Efron approximation (Prentice & Gloeckler, 1978; Singer & Willet, 2003).

However, when there are more tied observations than untied observations in the data, then none of the three tie-handling methods would work. Singer and Willet (2003) suggested that discrete-time methods should be used for such cases instead of continuous-time survival analysis. Similarly, Hess and Persson (2010) compared estimates from a Cox proportional model and a discrete model. The results showed that a Cox proportional model resulted in biased coefficients and standard errors from the data with heavy ties.

2.3. The mixed use of discrete or continuous models in the literature

To explore how survival analysis has been used in the field, thirteen studies that used survival analysis as the main statistical tool were summarized. Most of the studies indicated confusion over two estimation methods regardless of time metrics. Six studies among the thirteen adopted a continuous time estimation method by building a Cox regression model for relatively large time metrics without much explanation for their choice of the continuous model (Doyle, 2006; Kelly, 2004, Kirby, Berends & Naftel, 1999; Murphy, Gaughan, Hume, & Moore, 2010; Murtaugh, Burns & Schuster, 1999; Plank, DeLuca, & Estacion, 2008). Moreover, the seven studies that used a year as a time metric adopted either discrete (Donaldson & Johnson, 2010; Ma & Willms, 1999; Singer, 1992) or continuous survival analysis (Doyle, 2006; Kelly, 2004, Kirby, Berends, & Naftel, 1999; Murphy, Gaughan, Hume, & Moore, 2010). Four studies that chose a semester as a time metric used discrete (Calcagno, Crosta, Bailey & Jenkins 2007; Kahn & Schwalbe, 2010; Singer, Davidson, Graham & Davidson, 1998) or continuous analysis (Murtaugh, Burns & Schuster, 1999). Two studies used a month as a time metric for discrete (Jacobs & King, 2002) or continuous analysis (Plank, DeLuca, & Estacion, 2008). Most of all, these studies did not indicate a proper rationale for using either a continuous or a discrete model. Doyle (2006) chose a continuous-time survival analysis because of the common practice of building a continuous model based on the characteristics of the target events in the author's field of research.

Another problem of prior studies is the fact that the researchers adopted either a continuous or a discrete model without considering important factors for survival model-building. For example, Plank, DeLuca, and Estacion (2008) chose a continuous model with an exact method because the authors did not find any difference in the parameter estimates between continuous and discrete models without consideration of hazard rates. Obviously, a guideline that suggests the proper use of either a discrete or a continuous model is needed.

2.4. A logit model and a complementary log-log model

In the survival analysis, the risk (probability) of a target is estimated as a hazard ($h(t_{ij})$) as follows:

$$h(t_{ij}) = n\ events_j / n\ at\ risk_j$$

Where n $events_j$ refers to the number of samples that experience an event in time period j and n at $risk_j$ indicates the number of samples that have not yet experienced the event up to the time period j.

For the discrete hazard model, this study adopted a logit model; one of the most common discrete models (Allison, 2010; Singer & Willet, 2003). To estimate a hazard, a logit model uses the following link function:

$$Logit\ (p) = log[p_i/(1-p_i)]$$

Where p_i stands for the probability of an event for the ith observation. In the study, Allison's logit model (2010) was used to estimate the parameters for the discrete hazard model.

$$Logit\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \ldots + \alpha_J D_{Jij}$$

Where intercept parameters $\alpha_1, \alpha_2, \ldots, \alpha_J$ indicate the logit hazard of the respective time periods.

Following the recommendations of prior studies for the case of heavy ties when building a Cox proportional model, this study adopted a complementary log-log (clog-log) model for the continuous hazard model (Allison, 2010; Hosmer & Lameshow, 1999). The clog-log model was equivalent to the exact method of the Cox proportional model, which is known to generate the most precise parameter estimates in the case of strong ties. In particular, the clog-log model was suitable for this study because the clog-log model produced a clog-log hazard for each time period, which can be compared with those from the discrete model.

The clog-log model transforms a hazard into a complementary log-log probability. The logarithm of the negated logarithm of the probability of event nonoccurrence is as follows:

The hazard of clog-log = log(-log(1-probability))
The study adopted the clog-log hazard by Allison (2010) as follows:
Clog-log $h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \ldots + \alpha_J D_{Jij}$

Where intercepts $\alpha_1, \alpha_2, \ldots, \alpha_J$ indicate the clog-log hazard for each time period.

The specifications of the clog-log model are very similar to those of the discrete model. The only difference is that the clog-log model estimates a clog-log hazard while the discrete model estimates a logit hazard.

2.5. Existing studies: a logit vs. a clog-log model

Studies have been conducted to compare a logit model with a clog-log model and found no difference between them. Allison (2010) empirically investigated the difference between these two methods by using the same data and found similar parameter estimates. The p-values of covariate estimates of the two methods were similar, while parameter estimates of a logit model were slightly larger than those of a clog-log model. Allison argued that parameter estimates from a discrete model tend to be larger than those of a clog-log model. Similar results were also noted in other studies. Colosimo, Chalita, and Demétrio (2000) compared logit and clog-log models for data from various conditions of 12-time metrics, a sample size of 198, hazard rates between 0.005 to 0.44, and a 25% censoring proportion. Using likelihood ratio tests, the researchers did not detect discrepancies between these two models. Corrente, Chalita, and Moreira (2003) attempted to provide a guideline for a clog-log or a logit model using data for 286 samples with high-tied events during 52 intervals. The authors were not able to identify a discrepancy between the two models by using residuals and fit statistics.

In sum, the study findings on the difference between the two models are inconclusive. This is partly because these studies have yet to incorporate essential factors into their studies. In response to the findings, this study compared the two models taking important factors into account. The first factor that the study considered is a hazard. In addition, this study also paid attention to the factors that influence the hazard, such as time metric, censoring proportion, and sample size.

2.6. Hazards

The hazard has been identified as an important factor that determines the discrepancy between a logit and a clog-log model. According to Hosmer and Lameshow (1999), while the outcome of a clog-log model is similar to that of a logit model, when a hazard is smaller than 0.15, the difference between the two models becomes notable when a hazard is greater than 0.15. Similarly, Singer and Willet (2003) found a difference between a logit and a clog-log model outcomes during periods with a high hazard. However, the authors did not notice differences in parameter estimates from the two models. Hess (2009) also found a great difference in the estimated parameters from clog-log and logit models when a hazard was as high as 0.3. However, no difference was noted in the parameter estimates for the periods with a hazard lower than 0.3.

2.7. Time metrics

The major difference between discrete and continuous survival models is the unit used to measure the timing of an event. A discrete model uses comparatively large time metric, resulting in a small number of waves (frequency). A continuous model uses a continuous time metric, leading to a large number of waves. A small number of waves will force many events into the same wave even though the events occurred at different time points. Thus, the presence of many tied observations will lead to a high hazard rate. This will increase the potential discrepancy between the logit and clog-log models, while a large number of waves will lead to a low chance of a discrepancy, by having a low hazard rate.

The study conducted by Hofstede and Wedel (1999) showed the effect of the size of time metrics on the potential discrepancy between the logit and clog-log models. The researchers built both continuous and discrete-time hazard models using different sizes of time metrics. They found that when time metrics were aggregated into a large unit, the discrepancy in parameter estimates between continuous and discrete-time hazard models was large. In other words, when a small time metric (originally a day) was aggregated into a week or a month, the hazard estimates in discrete models were overestimated while those in continuous models were underestimated.

2.8. Censoring proportion

Censored data represents the sample data for which a target event does not occur during the data collection period. Censoring is an important factor when comparing discrete and continuous survival models. This factor is important because the censoring proportion influences the number of cases that experience the target event, which changes the number of tied observations and the hazard rates. Despite the importance of the censoring proportion in distinguishing between discrete and continuous models, there is little or no research that paid attention to the different proportions of censoring as a main interest.

Hertz-Picciotto and Rockhill (1997) discussed the importance of censoring in survival analysis, although the authors did not include the censoring proportion in their model. According to the authors, censoring reduced the estimate biases because it kept the tie proportion low. A study conducted by Colosimo, Chalita, and Demétrio (2000) generated data with censoring proportions of 0%, 30%, and 60% to examine the effects of censoring on the two methods. However, the researchers explored the effects of censoring proportions in a variety of sample sizes. Therefore, they did not solely explore the effect of the censoring proportion. Hess (2009) also paid attention to low (16%-20%) and high censoring proportions (24%-27%). Hess found that increasing the censoring proportions increased the standard deviations of the parameter estimates but did not result in a discrepancy in the parameter estimates.

2.9. Sample size

Given the fact that a large sample size leads to a more robust statistical analysis, it is surprising that there is a lack of research on sample size relating to the proper use of survival analysis. When research can be found, it is complicated by the mixed results regarding the effect of sample size for discrete and continuous hazards models. In Hess' study (2009), the logit and clog-log models displayed a difference when the sample size was equal to or more than 1,000. This result was due to the large sample having more tied observations. In contrast, the study conducted by Hertz-Picciotto and Rockhill (1997) showed that the discrepancy between the two models was larger with a small sample. The researchers compared three methods (the Breslow, Kalbflesch-Prentice, and Efron methods) of treating ties for the Cox proportional model by differentiating between sample sizes. Hertz-Picciotto and Rockhill used four different sample sizes of 50, 100, 500, and 1,000 and found that all three methods, treating ties, displayed biased estimates when the sample size was 50. However, the discrepancy was due to the design of the study. The researchers fixed the number of ties across different sample sizes, resulting in the creation of a high hazard when the sample was small. Therefore, their study emphasized the tied observations more than the influence of sample size. Unfortunately, there are few studies that have examined the effect of sample size while keeping other factors constant. This study sought to contribute information in regard to the sample size and its effect on survival model outcomes.

3. Methodology

In spite of the conceptual difference between discrete and continuous survival models, the current literature review reveals the mixed use of these two models without a clear rule. In response, this study compared two analyses – logit and clog-log survival analysis – in 60 conditions by combining three factors. The three factors that were combined are three time metrics (4, 12, and 48); five censoring proportions (0%, 20%, 40%, 60%, and 80%); and four sample sizes (50, 100, 500, and 1,000). The study simulated the data by combining varying levels of the factors. Using the simulated data, both discrete and continuous survival analyses were built to compare the parameter estimates and fit statistics.

3.1. Data generation

The data was generated using SAS with a variety of time metrics, censoring proportions, and sample sizes. These three factors combined determine hazards. , The number of cases that experienced the target event was determined for each data set was based on the total sample size and censoring proportions. The study generated 60 combinations of data when the three-time metrics, five censoring proportions, and four sample sizes were combined.

The study specified three types of time metrics of 4, 12, and 48 which emulated educational conditions. For example, for the case of 4 years of high school, one can choose the base unit as a year, quarter, or month. When a year is chosen, 4 is the time metric; when a quarter is chosen, 12 is the time metric; and when a month is chosen, 48 is the time metric.

The study included 5 censoring proportions to make a possible range of 0%, 20%, 40%, 60%, and 80%. The censoring proportion indicates cases that did not experience the target event. The study assumed the right-hand censoring, meaning that the censoring occurred at the end of the data collection period.

This study selected the following sample sizes: 50, 100, 500, and 1,000, as prior researchers such as Hertz-Picciotto and Rockhill (1997) used. The raw hazard was calculated by dividing the number of the risk set (cases that did not experience the target event until that time) with the number of events in each period. First, the total number of events (cases that experienced the target event) was obtained by multiplying the total sample size by an uncensored proportion (100%-censoring proportion). The number of events (tied observations) per period was calculated by dividing the total number of events by the number of periods. The hazard rate is estimated as a proportion of the number of tied observation (events) per each period out of the number of the risk set (the total number of cases in which the target event did not occur to up to that period). This study assumed right-hand censoring and no missing data.

3.2. The model of the study: Discrete (logit model) vs. Cox (exact; clog-log model)

The three equations for the logit and clog-log models were specified as follows:

Using 4-time metrics:

$$Logit\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \alpha_3 D_{3ij} + \alpha_4 D_{4ij} \quad (3.1)$$

$$Clog\text{-}Log\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \alpha_3 D_{3ij} + \alpha_4 D_{4ij} \quad (3.2)$$

Using 12-time metrics:

$$Logit\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \alpha_3 D_{3ij} + \ldots + \alpha_{10} D_{10ij} + \alpha_{11} D_{11ij} + \alpha_{12} D_{12ij} \quad (3.3)$$

$$Clog\text{-}Log\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \alpha_3 D_{3ij} + \ldots + \alpha_{10} D_{10ij} + \alpha_{11} D_{11ij} + \alpha_{12} D_{12ij} \quad (3.4)$$

Using 48-time metrics:

$$Logit\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \alpha_3 D_{3ij} + \ldots + \alpha_{46} D_{46ij} + \alpha_{47} D_{47ij} + \alpha_{48} D_{48ij} \quad (3.5)$$

$$Clog\text{-}Log\ h(t_{ij}) = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \alpha_3 D_{3ij} + \ldots + \alpha_{46} D_{46ij} + \alpha_{47} D_{47ij} + \alpha_{48} D_{48ij} \quad (3.6)$$

3.3. Maximum likelihood estimation

As an estimation method, this study adopted the maximum likelihood method. This method estimates population parameters by maximizing the probability that the sample data will be observed. The likelihood function stands for the likelihood of observing the pattern of event occurrence or non-occurrence in a dataset. In the case of the discrete (logit) model, the likelihood function was specified as follows:

$$\text{Likelihood} = \prod_{i=1}^{n} \prod_{j=1}^{Ji} h(t_{ij})^{EVENTij} (1 - h(t_{ij}))^{(1-EVENTij)}$$

Where $h(t_{ij})$ refers to the probability that the event will occur to an individual *i* in the *j* period. *EVENTij* indicates whether the event happens to an individual *i* in the *j* period; 1 indicates an event occurrence, while 0 indicates no event occurrence.

The likelihood function shown above was simplified into the following log-likelihood (LL) function.

$$LL = \sum_{i=1}^{n} \sum_{j=1}^{Ji} Event_{ij} \, log \, h(t_{ij}) + (1 - Event_{ij}) log \, (1 - h(t_{ij}))$$

(Singer & Willet, 2003)

The maximum likelihood estimation function for the clog-log model is as follows:

$$\text{LL} = \sum_{i=1}^{n} \sum_{j=1}^{Ji} \text{Event}_{ij} \log \left(1 - \exp\left(-\exp\, h(t_{ij})\right)\right) + (1 - \text{Event}_{ij}) \log \left(-\exp\left(-\exp\, h(t_{ij})\right)\right)$$

(Franklin, 2005)

3.4. Model comparison

After building two models, the study compared the parameter estimates with significance levels (the hazard of the respective period) and model fit statistics. The hazard estimates from the two models were also compared with significance levels of 0.05 and 0.01.

To assess fit for the two models, the study used the goodness-of-fit statistics for 60 conditions. The goodness-of-fit statistics used -2LL, which was converted from a log-likelihood statistic (LL). In particular, the study adopted the Akaike Information Criterion (AIC) because the models under comparison are non-nested models (see Allison, 2010; Singer & Willet, 2003). The smaller the values are, the better fit the model demonstrates. The AIC is calculated by being penalized based on the number of parameters as follows:

$$AIC = -2LL + 2p$$

Where *p* indicates the number of parameters.

4. Results

4.1. Hazard estimates

With regard to time metrics, the smaller time metrics were associated with more discrepancies as shown in Table. The use of four-time metrics revealed the most discrepancies. Among all 20 conditions of the 4-time metrics, 8 conditions displayed discrepancies between two models in terms of the significance levels of the estimates, between the two models. The use of 12-time metrics revealed discrepancies in 7 conditions, while the use of 48-time metrics showed discrepancies in only 2 conditions.

Table. Hazard Estimates across Different Time Metrics. Censoring Proportions, and Sample Sizes

Different Conditions		Logit Estimate	Logit SE	Clog-Log Estimate	Clog-Log SE	Discrepancy Frequency	AIC Difference
Time Metrics	4 Times	-1.700	0.090	-1.870	0.230	8	517
	12 Times	-2.926	2.098	-2.664	2.165	7	1,935
	48 Times	-4.039	0.596	-3.807	0.785	2	8,567
Censoring Proportions	0%	-1.682	21.699	-2.123	2.070	9	5,581
	0.2%	-2.292	7.355	-2.563	0.555	7	6,564
	0.4%	-2.916	0.573	-2.930	0.564	1	7,718
	0.6%	-3.192	0.566	-3.201	0.560	0	8,857
	0.8%	-3.672	1.154	-3.675	1.067	0	10,712
Sample Sizes	50	-2.575	16.334	-2.798	2.252	6	330
	100	-3.237	8.514	-3.392	1.323	4	742
	500	-3.674	2.947	-3.791	0.590	4	4,376
	1,000	-3.669	2.296	-3.789	0.419	3	8,739

The examination of censoring proportions revealed that fewer censoring proportions were associated with greater discrepancies between the two models, as provided in Table. The number of estimates that showed differences between the two models was the highest when the censoring was 0%. Out of 12 conditions of 0% censoring, the parameter estimates of 9 conditions revealed different significance levels. The three conditions that showed matched results from the two models were the data with large samples (500 and 1000) and large time metrics (48).

The non-matching results were displayed with a 20% censoring proportion, in which 7 out of 12 conditions produced non-matching estimate results. However, a smaller number of discrepancies was detected for the 40% censoring proportion; only one condition displayed a discrepancy. No discrepancies were detected for the 60% and 80% censoring proportions.

Among the four different sample sizes (50, 100, 500, and 1,000), the smaller samples showed more discrepancies as revealed in Table. Out of fifteen conditions that had 50 cases, six conditions revealed discrepancies. In the samples of 100 and 500 cases, four conditions showed discrepancies; while in the 1000-case samples, only three conditions showed discrepancies.

The study also calculated and compared the raw hazard rates of the parameter estimates that showed discrepancies between the clog-log and logit models. The results showed that the range of the hazard rates of the estimates ranged from 0.12 to 0.50, with 14 cases having hazard rates higher than 0.21. The important findings of the study regarding the hazard rates and discrepancies of the two models were that: when the parameter estimates showed a discrepancy, they displayed high hazard rates. However, the reverse was not true. In other words, the high hazard rates did not always lead to discrepancies.

Overall, the magnitude of the parameter estimates of the logit and clog-log models were similar, while the magnitudes of the logit model were larger than those of the clog model (Allison, 2010). The average hazard estimates from both the logit and clog-log models are included in Table. It is important to note that a big discrepancy in the hazard estimates was detected in the case of 0% censoring. The hazard estimates of the logit models ranged from 14.2029 to 17.2029, while those of the clog-log models ranged from 2.6824 to 2.8746.

4.2. AIC differences

The study found that the models of logit with a small number of time metrics, a small proportion of censoring, and a small sample tended to show a low AIC value (See Table for details). For example, the smallest AIC value (177) was found in the logit model with 0% censoring, 4 time metrics, and 50 cases. The largest AIC (86,831) was found in the clog-log model with 80% censoring, 48 time metrics, and 1,000 cases.

In all conditions, the logit models showed better fit statistics with lower values of AIC than the clog-log models for the same conditions (showing a difference ranging from 24 to 26,350). The smallest AIC difference (24) was from the comparison of a logit model (AIC =177) with a clog-log model (AIC=202) with 0% censoring, 4 time metrics, and 50 cases. The largest difference (26,350) resulted from the comparison of a logit model (AIC= 6,011) with a clog-log model (AIC= 86,361) with 80% censoring, 48 time metrics, and 1,000 cases.

5. Discussion

This study attempted to provide a guideline for empirical researchers of survival analysis in response to the confusion regarding the proper use of discrete or continuous models of survival analysis.. The study adopted a logit model as a discrete model and a clog-log model as a continuous model. Most importantly, this study confirmed the need for a guideline through the study's results, which indicated discrepancies between the discrete and continuous models in many conditions. The study also verified the importance of time metrics, censoring proportions, and sample size in addition to hazard rates in choosing survival models. Furthermore, the study addressed the interaction effects of the three factors affecting the discrepancy of the outcomes of a logit model and a clog-log model in the same condition.

To examine the effect of the three factors, this study generated 60 sets of data by combining different levels of the factors: time metrics (4, 12, and 48); censoring proportions (0%, 20%, 40%, 60%, and 80%); and sample size (50, 100, 500, and 1,000). After employing two methods to each of 60 simulation conditions, the study compared the parameter estimates and fit statistics to evaluate the performance of the two models.

Only a small number of time metrics were associated with greater discrepancies. This discrepancy pattern was particularly detected with small censoring proportions. This may have happened because, when there is a small number of time metrics, there will be a large number of tied observations and a high hazard rate for a period with a fixed number of cases. Thus, when the data is measured using large time metrics, such as 4 or 12 times, with small censoring proportions (less than or equal to 20%), it is recommended to build a discrete model regardless of sample size. For these conditions, a continuous model would produce biased estimates. When the data is measured in fine units such as 48 time metrics, the building of either continuous or discrete models should be considered.

The study results indicated that a small censoring proportion was associated with greater discrepancies because low censoring indicates more events during each time period. In particular, discrepancies that occurred in the conditions with less than 40% censoring. The discrepancies are more pronounced when low censoring was combined with small sample sizes and small-time metrics. Thus, this study recommends the following: when the censoring proportions are equal to or less than 40%, it is desirable to build a discrete model. It is highly recommended to build a discrete model when there is are small censoring proportions in small samples (equal to or less than 500) with small time metrics.

In regard to sample size, smaller samples were associated with greater discrepancies between the two models. However, compared with censoring proportions and time metrics, the sample size did not lead to much discrepancy between the two models. As suggested by previous studies, higher hazard rates were associated with greater discrepancies. However, there were some exceptions. Therefore, the study suggests that a choice of model based on hazard rates should be made in consideration of other three factors.

Overall, the discrete models showed better fit statistics than the continuous models. In all 60 conditions, the discrete models indicated comparatively small deviances. Singer and Willet (2003) suggested that discrete models should be used when there are more tied observations than unique observations.

By employing both continuous and discrete models to the simulated data, the study estimated hazard rates for the parameter estimates, exploring discrepancies between two models. The main focus of the study was on detecting the effects of the three factors of time metrics, censoring proportions, and sample size in addition to hazard rates. However, the study did not pay attention to other covariates (other predictor variables). The study suggests that future studies should extend the scope by including other important covariates and estimate hazards in order to examine interactional effects.

References

Allison, P. (2010). *Survival Analysis Using SAS®: A Practical Guide Second Edition*. Cary, NC: SAS Publishing.

Calcagno, J. C., Crosta, P., Bailey, T., & Jenkins, D. (2007). Does Age of Entrance Affect Community College Completion Probabilities? Evidence From a Discrete-Time Hazard Model. *Educational Evaluation and Policy Analysis*, 218-235.

Colosimo, E. A., Chalita, L. V. A. S., & Demétrio, C. G. B. (2000). Tests of Proportional Hazards and Proportional Odds Models for Grouped Survival Data. *Biometrics, 56*(4), 1233-1240.

Corrente, J. E., Chalita, L., & Moreira, J. A. (2003). Choosing between Cox proportional hazards and logistic models for interval- censored data via bootstrap. *Journal of Applied Statistics, 30*(1), 37 - 47.

Donaldson, M. L., & Johnson, S. M. (2010). The Price of Misassignment: The Role of Teaching Assignments in Teach For America Teachers' Exit From Low-Income Schools and the Teaching Profession. *Educational Evaluation and Policy Analysis 32*(2), 299-323.

Doyle, W. R. (2006). Adoption of Merit-Based Student Grant Programs: An Event History Analysis. *Educational Evaluation and Policy Analysis*, 259-285.

Hertz-Picciotto, I., & Rockhill, B. (1997). Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression. *Biometrics, 53*(3), 1151-1156.

Hess, W. (2009). A flexible hazard rate model for grouped duration data. *Working papers series (Tartu Ülikool. Majandusteaduskond)*.

Hess, W., & Persson, M. (2010). *The duration of trade revisited: Continuous-time vs. discrete-time hazards*.Unpublished manuscript.

Hofstede, F. t., & Wedel, M. (1999). A Monte Carlo study of time aggregation in continuous-time and discrete-time parametric hazard models. *Economics Letters, 58*(2), 149-156.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: regression modeling of time to event data*. New York, N.Y.: Wiley.

Jacobs, J. A., & King, R. B. (2002). Age and College Completion: A Life-History Analysis of Women Aged 15-44. *Sociology of Education, 75*(3), 211-230.

Kahn, J. M., & Schwalbe, C. (2010). The timing to and risk factors associated with child welfare system recidivism at two decision-making points. *Children and Youth Services Review, 32*(7), 1035-1044.

Kelly, S. (2004). An event history analysis of teacher attrition: Salary, teacher tracking, and socially disadvantaged schools. *Journal of Experimental Education, 72*(3), 195-220.

Kirby, S. N., Berends, M., & Naftel, S. (1999). Supply and Demand of Minority Teachers in Texas: Problems and Prospects. *Educational Evaluation and Policy Analysis*, 47-66.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis : Techniques for censored and truncated data*. New York: Springer.

Ma, X., & Willms, J. D. (1999). Dropping Out of Advanced Mathematics: How Much Do Students and Schools Contribute to the Problem? *Educational Evaluation and Policy Analysis*, 365-383.

Murphy, T. E., Gaughan, M., Hume, R., & Moore, S. G. (2010). College Graduation Rates for Minority Students in a Selective Technical University: Will Participation in a Summer Bridge Program Contribute to Success? *Educational Evaluation and Policy Analysis*, 70-83.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*(3), 355-371.

Plank, S. B., DeLuca, S., & Estacion, A. (2008). High School Dropout and the Role of Career and Technical Education: A Survival Analysis of Surviving High School. [Article]. *Sociology of Education, 81*(4), 345-370.

Prentice, R. L., & Gloeckler, L. A. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics, 34*(1), 57-67.

Singer, J. D. (1992). Are special educators' career paths special? Results from a 13-year longitudinal study. *Exceptional Children, 59*(3), 262-279.

Singer, J. D., Davidson, S. M., Graham, S., & Davidson, H. S. (1998). Physician Retention in Community and Migrant Health Centers: Who Stays and for How Long? *Medical Care, 36*(8), 1198-1213.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford; New York: Oxford University Press.