Contemporary Mathematics and Applications
Vol. 5, No. 2, 2023, pp. 79-87.

Survival Analysis and Hazard of Log Logistic Distribution on Type I Censored Data Parametrically

Ruth Hosana^1, Ni Wayan Widya Septia Sari^2, Ardi Kurniawan^3
^{1,2,3}Departemen of Matematics, Faculty of Science and Technology, Airlangga University
Jl. Dr. Ir. H. Soekarno, Surabaya, 60116, Indonesia

^{3}Corresponding author: ardi-k@fst.unair.ac.id

Abstract. Survival Analysis is a research method that examines the survival time of individuals or experimental units in relation to events such as death, disease, recovery, or other experiences. This study utilizes a parametric survival analysis model with a 2 parameter log logistic distribution and Maximum Likelihood Estimation (MLE) method to analyze the survival of students during their study period. The log logistic distribution is chosen due to its ability to capture early or late failure patterns. The objective of this research is to analyze type I censored survival data using the log logistic distribution applied to secondary data on student study duration. The dataset consists of 98 observations. The calculated values for the β and γ parameters of the 2 parameters log logistic distribution are 2.12831 and 0.0918891, respectively. The probability of students completing their studies by semester 8 (hazard function h(8)) is 0.370102, while the probability of students continuing their studies in semester 9 (survival function s(9)) is 0.320817.

Keywords: *Survival Function*, *Hazard Function*, *Log Logistic Distribution*, *Type I Censoring*, *Study Duration*.

1 Introduction

Survival analysis is an investigation into the survival time of an individual or experimental unit. The events studied in survival analysis are related to death, disease onset, recovery, or other experiences that occur to an individual [1]. According to Lawless (1982), parametric survival analysis models are widely used in medical research to estimate survival times [2]. These survival models assume specific parametric forms for the distribution of survival times, and one of the commonly used estimation methods is maximum likelihood estimation to estimate the distribution's parameter values [3].

One distinguishing factor of survival analysis from other statistical fields is the presence of censoring. Censored observations refer to survival time observations that are not precisely known and are limited by time or other reasons [4].

According to Brahmantyo et al. (2015), Type I censoring occurs when the experiment is terminated after reaching a predetermined time, simultaneously ending the participation of all n individuals in the experiment [5]. In Type I censoring, parametric analysis requires a goodness of fit test for the data distribution, such as the Anderson Darling test. The Anderson Darling test is a statistical test that utilizes specific distributions such as normal, lognormal, log logistic, exponential, Weibull, and logistic distributions [6][7].

The log logistic distribution is useful for describing the occurrence rate of events while considering survival until a specific time. Applying the log logistic distribution in survival analysis provides insights into the timing and probability of events. The log logistic distribution has two main parameters: a scale parameter that determines the data's size and dispersion, and a location parameter that affects the curvature of the distribution curve [8].

According to the Regulation of the Minister of Research, Technology, and Higher Education No. 44/2015, the maximum study duration allowed for undergraduate students is 7 years. Prolonged study duration can impact students' future careers and increase the financial costs of education. Triyanto and Sulistyorini (2019) define study duration as the total time required for students to complete their study programs, including academic leave, suspension, or changing majors [9]. Several factors contribute positively to faster completion, such as younger age at enrollment, strong ethnic backgrounds, and good academic achievements [10].

Previous research has been conducted on survival analysis of factors influencing students' study duration. Sari and Marfuah (2019) used the Cox Regression Proportional Hazard method with factors including age, gender, school background, student status, and class attendance [11]. Zhang and Lu (2015) conducted an additive hazard regression analysis for interval censored data with Weibull distribution [12]. Another study by Sun et al. (2018) introduced robust estimation for interval censored survival data using the exponential distribution [13]. The findings of Sun et al (2018) research suggest that robust estimation methods can handle data disturbances and provide more reliable estimates in the presence of extreme or unusual data.

Based on the aforementioned discussion, the authors aim to further develop research on survival analysis of students' study duration, specifically investigating the survival function and hazard function of the log logistic distribution in Type I censored cases. The parameter estimation method employed is the Maximum Likelihood Estimation (MLE), applied to the study duration data of students obtained from Fitriana (2016) [14].

2 Research Methodology

The data used in this study are secondary data obtained from Fitriana (2016) [14] regarding the study duration of students. The variables in this study include the time taken by students to complete their studies, with the following definitions for each variable:

a. t_i represents the study duration of students.
b. $d_i$is a censoring variable used to determine how long it takes for students to graduate with a bachelor's degree, categorized as follows:
   i. $d_i = 1$ indicates observed students.
   ii. $d_i = 0$ indicates censored students.

To achieve the objectives of this research, the following procedures were conducted:

1. Collecting information or data related to students' study duration.
2. Determining censored and observed data based on the obtained secondary data. Data is considered censored if the students' study duration exceeds 8 (eight) semesters.
3. Creating tabulations of interval data for the study duration and sorting them from the smallest to the largest.
4. Analyzing the goodness of fit of the data distribution using the Anderson Darling method.
5. Estimating the parameters of Type I censored data using the Log Logistic distribution through maximum likelihood estimation (MLE).
6. Applying step 5 to the Type I censored data to obtain the survival function.
7. Utilizing the obtained survival function on the Type I censored data to derive the hazard function.
8. Drawing conclusions based on the research objectives from the obtained results.

3 Results and Discussion

The length of student study data (in semesters) obtained was 98 data with a distribution of 69 data included in the uncensored data and 29 data included in the censored data [14]. According to Sudjana (2010), the data distribution fit test was used to determine whether the distribution of data in this study followed a certain distribution and could be analyzed parametrically [15]. Selection of the appropriate distribution model and accurate parameter estimation methods are key factors to obtain meaningful results in parametric survival analysis [16]. One of the methods used to see a distribution following a certain distribution is by looking at the Anderson Darling value. The results of the data distribution fit test with the Anderson Darling method are as shown in Table 1 below:

Table 1 Conformity od Data Distribution

Distribution	Anderson Darling (adj)
Loglogistic	**163,543**
Logistic	163,832
Lognormal	164,370
Exponential	169,650
Normal	164,813

Based on Table 1, it can be seen that the smallest Anderson Darling value is the Log Logistic distribution with 2 parameters. So that it can be interpreted, that the distribution of data in this study follows the Log Logistic distribution with 2 parameters. Furthermore, the estimation of the distribution parameters of the type I censored Log Logistic used in this study will be searched using the Maximum Likelihood (MLE) method with the following steps [17]:

1. Determine the general form of probability density function (PDF) with a log logistic distribution:

$$f(t,\gamma,\beta)=\frac{\gamma\beta t^{\beta-1}}{[1+\gamma t^{\beta}]^{2}} \qquad (1)$$

With parameters γ > 0, β > 0, and time variable t ≥ 0.
Based on the definition of the cumulative distribution function, the equation (1) is integrated from 0 to t and is obtained:

$$F(t,\gamma,\beta)=\int_0^t f(w,\gamma,\beta)\,dw=\int_0^t \frac{\gamma\beta w^{\beta-1}}{[1+\gamma w^{\beta}]^{2}}dw=\frac{\gamma t^{\beta}}{1+\gamma t^{\beta}} \qquad (2)$$

The general form of the survival function with log logistic distribution:

$$S(t,\gamma,\beta)=1-F(t,\gamma,\beta)=1-\frac{\gamma t^{\beta}}{1+\gamma t^{\beta}}=\frac{1}{1+\gamma t^{\beta}} \qquad (3)$$

2. Determine the type I censored likelihood function of log logistic distribution.
The type I censored likelihood function, which is denoted by $L(\gamma,\beta)$, is found according to the formula:

$$L(\gamma,\beta)=\prod_{i=1}^{r} f(t_i,\gamma,\beta)\prod_{i=r+1}^{n} S(t_i^{*},\gamma,\beta)$$

Using (1), (2) and (3) obtained:

$$L(\gamma,\beta)=\prod_{i=1}^{r}\frac{\gamma\beta t_i^{\beta-1}}{\left[1+\gamma t_i^{\beta}\right]^{2}}\cdot\prod_{i=r+1}^{n}\frac{1}{1+\gamma t_i^{*\beta}} \qquad (4)$$

3. Taking the natural logarithm (ln) of the likelihood function (4) and denoting it $L^{L}(\gamma,\beta)$, then:

$$L^{L}(\gamma,\beta)=ln\,L\,(\gamma,\beta)=\ ln\left(\prod_{i=1}^{r} f\,(t_i,\gamma,\beta)\prod_{i=r+1}^{n} S(t_i^{*},\gamma,\beta)\right)$$

Therefore,

$$\ln L\,(\gamma,\beta)=\ ln\left(\prod_{i=1}^{r}\frac{\gamma\beta t_i^{\beta-1}}{[1+\gamma t_i^{\beta}]^{2}}\right)+ln\left(\prod_{i=r+1}^{n}\frac{1}{1+\gamma t_i^{*\beta}}\right) \qquad (5)$$

or

$$\ln L\,(\gamma,\beta)=r\,(\ln\gamma+\ln\beta)+(\beta-1)\sum_{i=1}^{r}\ln(t_i)-2\sum_{i=1}^{r}\ln\left(1+\gamma t_i^{\beta}\right)-\sum_{i=r+1}^{n}\ln\left(1+\gamma t_i^{*\beta}\right) \qquad (6)$$

4. Using the MLE method to find the estimator values of the parameters that maximize the likelihood function in equations (5) and (6). The first step to get the estimator value that maximizes equations (5) and (6) is to reduce the equation to γ and β, then the results are equated to zero, so we get

$$\frac{\partial L_L(\gamma,\beta)}{\partial \gamma} = \frac{r}{\gamma}\ 2\sum_{i=1}^{r} \frac{t_i^{\beta}}{1+\gamma t_i^{\beta}} - \sum_{1=r+1}^{n} \frac{t_i^{*\beta}}{(1+\gamma t_i^{*\beta})} = 0, \text{ and} \tag{7}$$

$$\frac{\partial L_L(\gamma,\beta)}{\partial \beta} = \frac{r}{\beta} + \sum_{i=1}^{r} \ln{(t_i)} - 2\gamma \sum_{i=1}^{r} \frac{t_i^{\beta} \ln{(t_i)}}{1+\gamma t_i^{\beta}} - \gamma \sum_{1=r+1}^{n} \frac{t_i^{*\beta} \ln(t_i^{*})}{(1+\gamma t_i^{*\beta})} \tag{8}$$

Equations (7) and (8) are not explicit equations so the solution cannot be done directly, but the solution must be done numerically.

Application to Research Data

An example of implementing type I censored cases with a 2 parameter log logistic distribution will use the study time data provided by Fitriana (2016) [14].

a. Parameter estimation on both parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

Based on calculations, the result of estimating parameters β (*location*) and γ (*scale*) about study time data from Fitriana (2016) [14] can be seen in Table 2 below:

Tabel 2. Parameter Estimation of Log Logistic Distribution 2 Parameters

		Standard	95,0% Normal CI	
Parameter	**Estimate**	**Error**	**Lower**	**Upper**
Location	2,12831	0,0163518	2,09626	2,16036
Scale	0,0918891	0,0099619	0,0742991	0,113644

From Table 2 above, it can be seen that the estimators for the parameters β and γ are β=2,12831 and γ=0,0918891. Then from Table 2 it can be seen also the value of the confidence interval and standard error of each parameter. Thus, the form of the pdf distribution of the length of study time is:

$$f(t,\gamma,\beta) = \frac{1{,}955685 * t^{1{,}12831}}{[1 + 0{,}0918891 * t^{2{,}12831}]^2}$$

b. Log Logistic Distributed Data Survival Function

The form of estimation of the survival function on the length of student study is:

$$S(t,\gamma,\beta) = \frac{1}{1 + 0{,}0918891 * t^{2{,}12831}}$$

The results of the probability estimation of the survival function for the length of student study are given in Table 3 below

Table 3. Estimation of Student Survival Function

		95,0% Normal CI	
Time	**Probability**	**Lower**	**Upper**
7	0,879209	0,819565	0,921036
8	0,629898	0,548478	0,704544
9	0,320817	0,236894	0,418180
10	0,130491	0,076225	0,214424
11	0,050505	0,023568	0,104921
12	0,020218	0,007716	0,051916

From the output above, it is known that the probability of students continuing their studies in semester 7 is 0.879209; the probability that students are still continuing their studies in semester 8 is 0.629898; the probability that students are still continuing their studies in semester 9 is 0.320817; the probability that students are still continuing their studies in semester 10 is 0.130491; the probability that students are still continuing their studies in semester 11 is 0.050505; and the chance that students will continue their studies in semester 12 is 0.020218. Table 3 survival values can be presented in graphical form as shown in Figure 1.

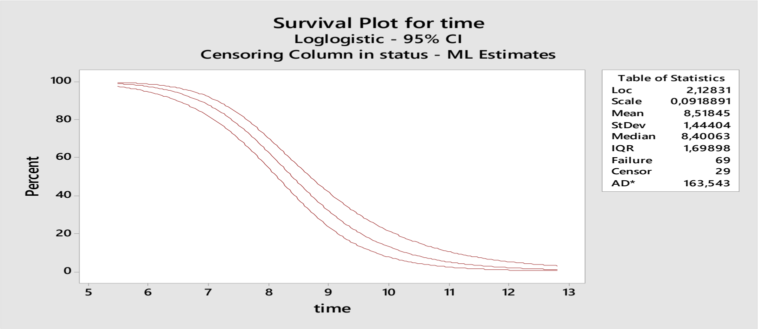

Figure 1. Survival Plot

c. **Log Logistic Distributed Hazard Data Function**

The hazard function of data with a log logistic distribution with parameters γ and β can be obtained by dividing the probability density function by the survival function, so we get

$$h(t,\gamma,\beta) = \frac{f(t,\gamma,\beta)}{S(t,\gamma,\beta)} = \left(\frac{\gamma\beta t^{\beta-1}}{\left[1+\gamma t^{\beta}\right]^{2}}\right) \Big/ \left(\frac{1}{1+\gamma t^{\beta}}\right) = \frac{\gamma\beta t^{\beta-1}}{1+\gamma t^{\beta}}$$

Thus, the hazard function h(t, γ, β) is obtained:

$$h(t,\gamma,\beta) = \frac{1{,}955685 * t^{1{,}12831}}{1 + 0{,}0918891 * t^{2{,}12831}}$$

Based on the calculation results, the estimated value of the hazard function is obtained as follows

Tabel 4. Estimation of Student Hazard Function

		95,0% Normal CI	
Time	**Probability**	**Lower**	**Upper**
7	0,120791	0,078964	0,180435
8	0,370102	0,295456	0,451522
9	0,679183	0,581820	0,763106
10	0,869509	0,785576	0,923775
11	0,949495	0,895079	0,976432
12	0,979782	0,948084	0,992284

From the output above, it is known that the probability that students have completed their studies in semester 7 is 0.120791; the probability that students have completed their studies in semester 8 is 0.370102; the probability that students have completed their studies in semester 9 is 0.679183; the probability that students have completed their studies in semester 10 is 0.869509; the probability that students have completed their studies in semester 11 is 0.949495; and the probability that students have completed their studies in semester 12 is 0.979782. From Table 4 the value of the hazard function can be presented in graphical form as shown in Figure 2.

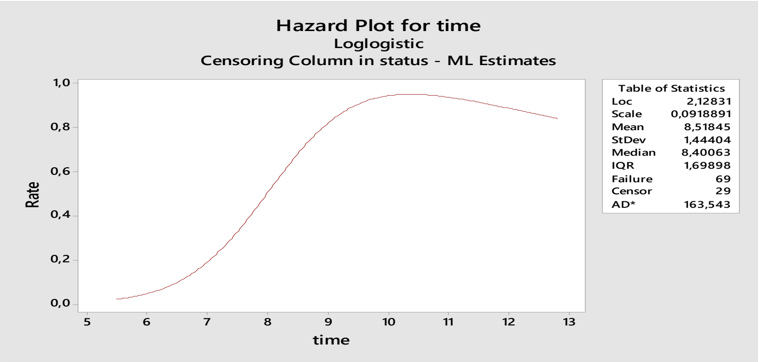

Figure 2. Hazard Plot

4 Conclusion

Based on the research conducted, the following conclusions can be drawn:

a. Of the 98 data, there are 29 students who are still continuing their censored (live) studies with 9 semesters of study with a survival chance of 0.320817 or 32% to 12 semesters with a survival chance of 0.020218 or 2%.
b. The distribution of the data in this study follows the Log Logistic distribution with 2 parameters where the values β=2.12831 and γ=0.0918891.
c. The probability that students have completed their studies up to semester 8 is 0.370102

5 References

[1] Kleinbaum, D.G. & Klein, M., 2012, *Survival Analysis: A Self Learning Text*, Springer Science & Business Media.

[2] Lawless, J.F., 1982, *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons.

[3] Lambert, P., 2017, *Parametric Survival Analysis in Medical Research: A Review*, Journal of Medical Devices, **11**(2), 020801.

[4] Lee, E.T. & Wang, J., 2003, *Statistical Methods for Survival Data Analysis*, John Wiley & Sons.

[5] Brahmantyo, A., Nur, M. & Sari, R., 2015, *Survival Analysis with Cox Regression Method*, Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Indonesia.

[6] Anderson, T.W. & Darling, D.A., 1954, *A Test of Goodness of Fit*, Journal of the American Statistical Association, **49**(268), 765-769.

[7] Law, A.M. & Kelton, W.D., 1991, *Simulation Modeling and Analysis*, McGraw Hill.

[8] Al Bahadili, H., 2015, *Estimation of the log logistic distribution parameters using multiple imputations*, Journal of Statistical Computation and Simulation, **85**(5), 1039-1052.

[9] Triyanto & Sulistyorini, 2019, *Analysis of On Time Graduation of Students Using Survival Analysis at Yogyakarta State University*, Journal of Universitas Negeri Semarang, **10**(2), 123-145.

[10] Terrell, M. & Martin, A., 2018, *Factors Influencing Time to Degree Completion Among College Students*, Journal of College Student Retention: Research, Theory & Practice, **20**(3), 277-294.

[11] Sari, R. & Marfuah, M., 2019, *Analysis of Survival and Factors Affecting Students's Study Duration*, Journal of Educational Science Research, **3**(2), 123-137.

[12] Zhang, L. & Lu, W., 2015, *Additive hazards regression analysis with interval censored data*, Computational Statistics & Data Analysis, **83**, 77-91.

[13] Shi, P. & Sun, J., 2018, *A robust estimator for interval censored exponential survival data*, Statistics in Medicine, **37**(26), 3801-3814.

[14] Fitriana, 2016, *Survival Analysis of Factors Affecting Students's Study Duration in Mathematics Education Batch 2010 using Cox Proportional Hazard Regression Method*, Undergraduate Thesis, Universitas Negeri Semarang.

[15] Sudjana, 2010, *Statistical Methods*. Bandung: Tarsito.

[16] Gupta, R. & Patel, K., 2019, *A Comprehensive Review of Parametric Survival Analysis Techniques*, Statistics in Medicine, **28**(14), 1989-2010.

[17] Smith, A., Johnson, B. & Williams, C., 2018, *Log logistic distribution as a suitable choice for survival analysis in the presence of time varying risks*, Journal of Statistics and Probability, **10**(3), 123-135.