CAUCHY – JURNAL MATEMATIKA MURNI DAN APLIKASI
Volume 5(2) (2018), Pages 55-63
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Fitting The First Birth Interval in Indonesia Using Weibull Proportional Hazards Model

Alfensi Faruk[1], Endro Setyo Cahyono[2], Ning Eliyati[3]

[1,2,3]Department of Mathematics, Faculty of Mathematics and Natural Science, Sriwijaya University

Email: alfensifaruk@unsri.ac.id, endrosetyo_c@yahoo.co.id, ningeliyati@gmail.com

ABSTRACT

The first birth interval (FBI) is defined as the duration of time spent by married couples to have their first child since the day of marriage. It is one of the indicators of women's fertility rate. The long FBI indicates the low fertility rate. Therefore, many governments of populous countries, such as Indonesia, attempted to prolong the FBI in order to prevent their country from overpopulation. In order to control the factors related to the FBI, it is important to determine the most significant factors and to quantify the effects of the significant factors on the FBI data. These tasks can be conducted by using survival analysis techniques. In particular, this research proposes Weibull proportional hazards (PH) model to analyze several socioeconomic and demographic factors, which may affect the FBI data in Indonesia. The data were obtained from 2012 Indonesian Demographic and Health Survey (IDHS) and consisted of 27488 ever married women aged 15-49 at the time of interview. In addition to the Weibull PH model, the data were also analyzed by using the other approaches, namely Kaplan-Meier (KM) method, logrank test, and PH model. The results indicated that the Weibull PH model performs very well compared to PH model. Moreover, the fitting of the Weibull PH model showed that age at the first birth, place of residence, women's educational level, husband's educational level, contraceptive knowledge, wealth index, and employment status had the significant effects on the FBI data in Indonesia.

Keywords: First birth interval, fertility rate, Weibull PH model

INTRODUCTION

Indonesia is the most densely populated country in South East Asia. According to Statistics Indonesia [1], there are about 259 million people of Indonesia in 2016. This has made Indonesia in the fourth rank as the world's most populous country behind China, India, and the United States. However, the huge amount of the Indonesian population is not along with the national economic condition which is still categorized as one of the third world countries. These facts show that the population density is still the main problem in Indonesia and need to be resolved.

Fertility is one of the factors that influence the fluctuation of the number of population. One of the indicators of fertility rate is the total fertility rate (TFR), which can be defined as the average number of children that would be born to a woman over her reproductive age. According to Islam [2], TFR can be reduced by increasing the age at marriage. However, this strategy is difficult to apply in the developed countries such as Indonesia due to the influence of social and cultural factors. Another alternative strategy is by controlling the FBI, which is defined as the time interval of a married woman to give birth her first child since the time of first marriage. If the FBI can be controlled, the next birth time is also automatically controlled [2].

Many works employed survival analysis to evaluate the factors that affect the FBI. It due to the fact that the FBI data commonly contain censored observations, that is the subjects who did not give birth until the end of the observation time. Consequently, the conventional statistical methods, such as multiple linear regression or logistic regression, are no longer appropriate to analyze the data set. Otherwise, survival analysis provides various statistical tools to handle any types of censored data [3]. One of the examples is the work by Hidayat et al. [4] which showed that the area of residence, education, and age were factors affecting the FBI in three Provinces in Indonesia. They applied the Cox model and Cox extended model in analyzing the FBI dataset. By using parametric models, Shayan et al. [5] showed that age at marriage, level of women's

Submitted: 20 Pebruary 2017 Reviewed: 31 December 2017 Accepted: 29 March 2018
DOI: http://dx.doi.org/10.18860/ca.v5i2.4060

education, and menstrual status had highly significant effects on the FBI in Pakistan. In most recent work, Faruk [6] concluded that age, women's educational level, and wealth index significantly affect the FBI in South Sumatra Province, Indonesia.

In this study, the Weibull PH model and PH model are considered to evaluate the socioeconomic and demographic factors that could significantly affect the FBI. Comparisons were made to find the best multivariate model for the FBI data in Indonesia. The Weibull PH model was used since the Weibull distribution is commonly fitted to survival data [7]. Meanwhile, the PH model was presented because this model is the most applied model in the time-to-event study [8]. In this research, the dataset contains the sample from all 33 Provinces in Indonesia. The data, then, were also analyzed by using KM method and logrank test in addition to Weibull PH model and Cox model. The data analysis in this research was supported by RStudio version 1.1.383.

The remainder of this paper is organized as follows: the next section will outline the materials and methods of this research. In section 3, both univariate and multivariate survival analysis techniques such as the KM method, the logrank test, the PH model, and the Weibull PH model are implemented to analyze the FBI data in Indonesia. Finally, the last section presents the main conclusions of this work.

MATERIALS AND METHODS

The FBI data were obtained from 2012 IDHS. Statistics Indonesia and United States Agency for International Development (USAID) conducted the survey from May to June 2012. They used stratified two-stage cluster sampling technique to determine the sample of the study. In particular, the cluster blocks, which are based on the Indonesia population census in 2010, are the first stage unit sampling. Meanwhile, the households in each cluster block are the second stage unit sampling. The surveyors used questionnaires to gain the information needed from the respondents. In addition, only the data from the complete questionnaires are included in this research.

The FBI (in months) is the survival time which was measured from the day of marriage up to the time of first give birth. Moreover, the time was considered as the variable of interest. In the analysis, eight covariates which consisted of one continuous covariate and seven categorical covariates were used in this work. These covariates were socioeconomic and demographic characteristics of the respondents based on the 2012 IDHS (Table 1).

In order to achieve the research objectives, this work consists of several steps. Firstly, the KM method was applied to the FBI data in order to estimate the survival and hazards functions. After that, the KM estimation results were plotted based on the groups in each categorical covariate. Then, the logrank test was applied to evaluate the differences between two or more survival functions in every categorical covariate. Next, the PH model and the Weibull PH model was fitted to the FBI data. Finally, the performance of the PH model and the Weibull PH was evaluated by using Akaike information criterion (AIC).

RESULTS AND DISCUSSION

3.1. Description of The Dataset

A total of 27448 ever married women aged 15-49 at the time of interview were included in this research. Of total of respondents, there were 25564 women who gave birth and 1884 women who did not give birth until the end of observation time. The latter were categorized as the right censored survival data. The summary of the frequency based on the characteristics and the status of the survival time can be seen in Table 1.

Table 1. Descriptive Statistics of Covariates and Their Survival Time Status

Covariate	Categories	Frequency	Status	
			#Censored	#Event
Age at the first birth	-	27448	1884	25564

Place of Residence	Rural	14522	944	13578
	Urban	12926	940	11986
Women's Educational Level	No Education	1121	93	1028
	Primary	10238	519	9719
	Secondary & above	16089	1272	14817
Husband's Educational Level	No Education	752	71	681
	Primary	9733	520	9213
	Secondary & above	16963	1293	15670
Contraceptive Knowledge	No	435	114	321
	Yes	27013	1770	25243
Access to Mass Media	No	2630	203	2427
	Yes	24818	1681	23137
Wealth Index	Poor	12388	893	11495
	Middle & above	15060	991	14069
Employment Status	No	10984	730	10254
	Yes	16464	1154	15310

3.2. The Estimation of Survival Function by Using KM method

Let T be a continuous random variable of survival times and $t \geq 0$ is the realization of T, then the probability density function of T is

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t<T<t+\Delta t)}{\Delta t}, \quad (1)$$

where $f(t)$ is a nonnegative function and also known as the unconditional failure rate [9]. The survival function, denoted by $S(t)$, is defined as the probability of the occurrence of an event at more than time t and can be formulated by

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(x)dx = 1 - F(t), \quad (2)$$

where $F(t)$ is a cumulative distribution function at time t. $S(t)$ is a nondecreasing function and has the properties $S(t) = 1$ at $t = 0$ and $S(t) = 0$ at $t = \infty$.

To estimate survival function $S(t)$ by using KM method, the n survival times (include censored time) were firstly arranged in order of increasing magnitude such that $t_1 \leq t_2 \leq \cdots \leq t_n$. Then, the survival function can be estimated using the formula

$$\hat{S}(t) = \prod_{t_r \leq t} \frac{n-r}{n-r+1}, \quad (3)$$

where t_r is uncensored time for $r \in \{1,2,\ldots,n\}$, and $n-r$ is the number of individuals in the sample surviving longer than time t_r.

In this work, the KM estimate of the survival function, $\hat{S}(t)$, was estimated by using Equation (3) and conducted by using RStudio. The estimation results for each group in categorical covariates of the FBI dataset in Indonesia were plotted in Figure 1.

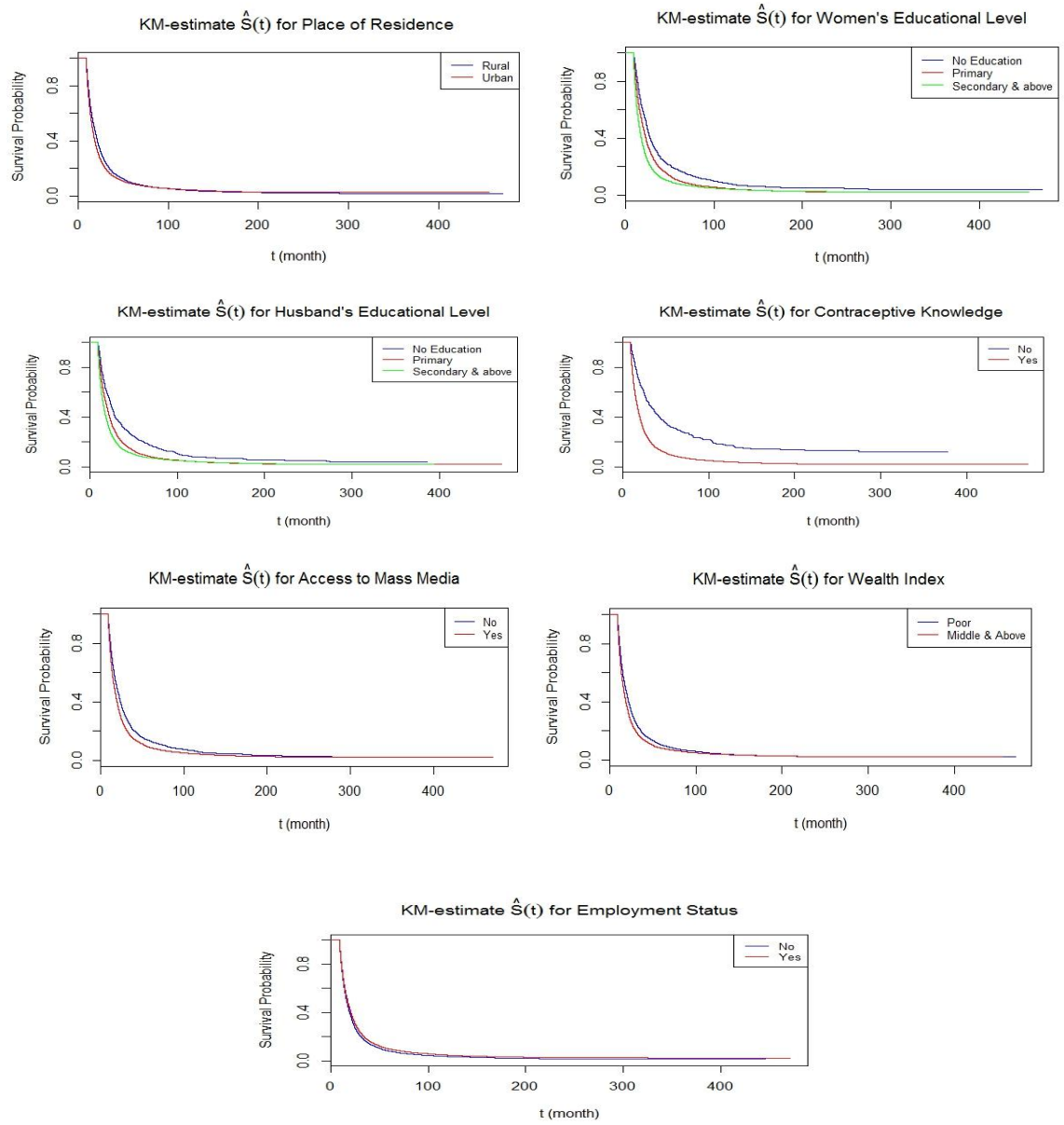

Figure 1. KM plots for survival of FBI in Indonesia

According to the plots in Figure 1, the obvious difference between survival curves can be seen in the survival curves which based on women's educational level, husband's educational level, contraceptive knowledge, access to mass media, and wealth index. Meanwhile, the difference of the survival curves which based on the other covariates such as place of residence and employment status is not obvious due to the estimate curves are very close each other. Therefore, a statistical test is needed to obtain more objective conclusion about the difference among the survival curves. In this work, the logrank test is applied to test the difference between two or more survival curves.

3.3. Comparing Survival Functions by Using The Logrank Test

In addition to KM plot, a statistical test is needed to obtain more convincing conclusion about the difference between the survival curves in each categorical covariates. For comparing two survival groups, the null hypothesis is $H_0 : S_1(t) = S_2(t)$ (no difference between two survival curves) with the test statistic

$$\frac{(O_2-E_2)^2}{Var\,(O_2-E_2)} \sim \chi^2_{df;\alpha}, \qquad (4)$$

where O_2 is the total number of failures in group 2, E_2 is the expected number of failures in group 2, df is degree of freedom which equals to 1, and α is level of significance. The null hypothesis (H_0) will be rejected if the $\chi^2 > \chi^2_{1;\alpha}$ or p-value < α.

To compare the survival curves for more than two groups, the null hypothesis is $H_0 : S_1(t) = S_2(t) = \cdots = S_n(t)$ (survival curves are similar) and the test statistic is [10]

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df;\alpha}, \quad (5)$$

where O_i is the total number of failures in group i, E_i is the expected number of failures in the group i, n is the total number of groups, and df is degree of freedom which equal to $n - 1$. If the $\chi^2 > \chi^2_{n-1;\alpha}$ or p-value < α, H_0 will be rejected. By using RStudio with 5% level of significance, the results of the logrank test between two or more survival curves of the FBI dataset in Indonesia are summarized in Table 2.

Table 2. Results of The Logrank Tests

Covariate	**Categories**	**df**	χ^2	$\chi^2_{df;0,05}$	**P-value**	**Conclusion**
Place of Residence	Rural Urban	1	80	3.481	0	Reject H_0
Women's Educational Level	No Education Primary Secondary &	2	48	5.591	0	Reject H_0
Husband's Educational Level	No Education Primary Secondary &	2	32	5.591	0	Reject H_0
Contraceptive Knowledge	No Yes	1	18	3.481	0	Reject H_0
Access to Mass Media	No Yes	1	90	3.481	0	Reject H_0
Wealth Index	Poor Middle &	1	16	3.481	0	Reject H_0
Employment Status	No Yes	1	46	3.481	8.27e-12	Reject H_0

As shown in Table 2, the χ^2 scores are more than $\chi^2_{df;0.05}$ and p–values for all logrank tests of each categorical covariate are less than the level of significance (α) = 0.05. Therefore, the null hypothesis H_0 can be rejected at the 5% level of significance. In other words, the difference of survival curves between categories for every categorical covariate based on the logrank test is statistically significant. The results are quite different with the work by Gebeyehu [11], which concluded that the difference between survival curves based on access to mass media was insignificant.

3.4. Fitting PH Model

PH model is represented by the relationship of the hazard function, the baseline hazard function, and one or more covariates in the form

$$h(t) = h_0(t) \exp(\boldsymbol{\beta}^t \mathbf{X}), \quad (6)$$

where $h(t)$ is the hazard function, $h_0(t)$ is the baseline hazard function which is left unspecified, $\boldsymbol{\beta}$ is a column vector of the regression coefficients, and $\mathbf{X}$ is a column vector of the covariates. The PH model assumes that there is proportionality of the hazard rate between any two individuals in the population.

The hazard ratio is defined as the ratio of the hazard functions for two subjects with different values of covariate X_1 and X_2. The formula of the hazard ratio is

$$H(t) = \frac{h_0(t) \exp(\beta_2 X_2)}{h_0(t) \exp(\beta_1 X_1)} = \frac{\exp(\beta_2 X_2)}{\exp(\beta_1 X_1)} = e^{\beta'(X_2 - X_1)}. \quad (7)$$

It can be seen that the hazard ratio in Equation (7) is independent of time. In other words, the hazard ratio for two individuals is constant over time. To estimate the regression coefficients of PH model, the partial likelihood method, which uses the partial likelihood function instead of the likelihood function as in the maximum likelihood estimation (MLE) method [12], should be used. The results of fitting the PH model to the FBI data in Indonesia in case of full model can be seen in Table 3. This fitting tests the hypothesis $H_0 : \hat{\beta}_i = 0, \forall i = 1 \dots n$ (covariate X_i is not statistically significant). The hypothesis H_0 is rejected if p-value $<$ level of significance (α).

Table 3. Results of Fitting The PH Model

Covariate	Category	$Exp\left(\widehat{\boldsymbol{\beta}}\right)$	P-value
Age at the first birth	-	1.004	0.005 [a]
Place of Residence	Rural [b]	-	
	Urban	0.993	0.604
Women's Educational Level	No Education [b]	-	
	Primary	1.079	0.032 [a]
	Secondary & above	1.267	1.97e-10 [a]
Husband's Educational Level	No Education [b]	-	
	Primary	1.151	0.001 [a]
	Secondary & above	1.224	3.14e-06 [a]
Contraceptive Knowledge	No [b]	-	
	Yes	1.706	< 2e-16 [a]
Access to Mass Media	No [b]	-	
	Yes	1.017	0.463
Wealth Index	Poor [b]	-	
	Middle & above	1.045	0.002 [a]
Employment Status	No [b]	-	
	Yes	0.938	5.53e-07 [a]

[a]*: significant at 10% level* , [b]*: reference category*

In the results of fitting PH model, as shown in Table 3, place of residence and access to mass media are two covariates which do not significantly affect the survival time at 10% level. Meanwhile, the rest of covariates have the significant effects on the FBI data in Indonesia at 10% level.

3.5. Fitting Weibull Proportional Hazards Model

In Cox PH model, the baseline hazards ($h_0(t)$) is left unspecified. However, if $h_0(t)$ is assumed to follow a certain survivor distribution, the model is called as parametric PH model. This model still holds the PH assumption as in the Cox PH, but it is a full parametric model instead of semiparametric. To estimate the unknown parameters, maximum likelihood estimation (MLE) method can be applied.

Among the survivor distributions, Weibull distribution is the most popular distribution. It because the survival and hazards functions of the Weibull distribution have the closed forms [13]. The PH model with a Weibull baseline distribution is given by

$$h(t) = \frac{p}{\lambda}\left(\frac{t}{\lambda}\right)^{p-1} \exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{X}), \qquad (8)$$

where $h(t)$ is the hazard function at time $t > 0$, p is shape parameter, λ is scale parameter, $\boldsymbol{\beta}$ is a column vector of the regression coefficients, and $\mathbf{X}$ is a column vector of the covariates. In RStudio,

the estimation of Equation (8) can be employed by using 'eha' package and phreg() function. The results of fitting the Equation (8) to the FBI data are summarized in Table 4.

Table 4. Results of Fitting The Weibull PH Model

Covariate	Category	$Exp\left(\widehat{\boldsymbol{\beta}}\right)$	P-value
Age at the first birth	-	1.002	0.063 [a]
Place of Residence	Rural [b]		
	Urban	0.958	0.002 [a]
Women's Educational Level	No Education [b]		
	Primary	1.145	0.000 [a]
	Secondary & above	1.385	0.000 [a]
Husband's Educational Level	No Education [b]		
	Primary	1.259	0.000 [a]
	Secondary & above	1.358	0.000 [a]
Contraceptive Knowledge	No [b]		
	Yes	2.054	0.000 [a]
Access to Mass Media	No [b]		
	Yes	1.023	0.329
Wealth Index	Poor [b]		
	Middle & above	1.030	0.039 [a]
Employment Status	No [b]		
	Yes	0.901	0.000 [a]

[a]*:significant at 10% level ,* [b]*:reference category*

According to Table 4, access to mass media is insignificant at 10% level and along with the results of PH model (Table 3). Meanwhile, the significance of place of residence is not similar to the results in Table 3, which concluded that place of residence is insignificant at 10% level. In addition, the results in Table 4 are also not along with the results of the work by [6] although it is consistent with the results of [11].

In order to assess the performance of PH model and Weibull PH model, it is common to use the AIC value. A regression model is said the better model if it has the lower AIC value than the other models. By using 'suvival' package with step() function in RStudio, it is obtained that the AIC values of PH model and Weibull PH model for the last step (best model) of backward stepwise procedure are equal to 477946.9 and 221723.4, respectively. It means that the performance of Weibull PH model is better than the PH model in fitting the FBI data in Indonesia. The best model of Weibull PH model based on the variable selection procedure is given in Table 5.

Table 5. The Best Model of Weibull PH Model Using Backward Stepwise Procedure

Covariate	Category	$Exp\left(\widehat{\boldsymbol{\beta}}\right)$	P-value
Age at the first birth	-	1.002	0.065 [a]
Place of Residence	Rural [b]		
	Urban	0.959	0.003 [a]
Women's Educational Level	No Education [b]		
	Primary	1.149	0.000 [a]
	Secondary & above	1.392	0.000 [a]

Husband's Educational Level	No Education [b]		
	Primary	1.263	0.000 [a]
	Secondary & above	1.362	0.000 [a]
Contraceptive Knowledge	No [b]		
	Yes	2.069	0.000 [a]
Wealth Index	Poor [b]		
	Middle & above	1.033	0.026 [a]
Employment Status	No [b]		
	Yes	0.901	0.000 [a]

[a]*:significant at 10% level ,* [b]*:reference category*

All covariates, that included in the best model in Table 5, are already significant at 10% level. The significant covariates based on the backward stepwise are similar with the results in Table 4. The clear difference between the models in Table 4 and Table 5 is that the best model excludes access to mass media from the best model. The exclusion of such covariate decreases the AIC value of the model in the previous step of the backward elimination procedure.

CONCLUSIONS

Although the KM method and logrank test are commonly used in the applications of survival analysis, they cannot account the effects of multiple covariates on the survival time simultaneously. Therefore, a typical survival regression model is needed to overcome such drawback. In this work, the PH model and Weibull PH model were applied to determine and quantify the effects of several socioeconomic and demographic factors on the FBI data in Indonesia. The model comparison results showed that the Weibull PH model was better than the PH model to fit the data. Moreover, the fitting of Weibull PH model by using backward stepwise procedure concluded that age at the first birth, place of residence, women's educational level, husband's educational level, contraceptive knowledge, wealth index, and employment status had the significant effects on the FBI data in Indonesia.

In this article, the PH assumption was assumed to the data. However, this assumption was not checked yet. Further research should account for this by using a graphical method or goodness of fit tests. If the data do not hold the PH assumption, further research also should consider the alternative models, such as accelerated failure time (aft) model or Cox extended model. Another extension is also can be made by also accounting the individual heterogeneity in the data by using frailty model.

ACKNOWLEDGMENTS

This project was sponsored by Sriwijaya University through Penelitian Sains Teknologi dan Seni (SATEKS) 2017. The authors are thankful to United States Agency for International Development (USAID) for contributing materials for use in the project.

REFERENCES

[1] Statistics Indonesia, Statistik Indonesia 2016 Statistical Yearbook of Indonesia 2016. Statistics Indonesia, 2016.

[2] M. S. Islam, "Differential determinants of birth spacing since marriage to first live birth in rural Bangladesh," *Pertanika Journal of Social Science and Humanities*, vol. 17, no. 1, pp. 1–6, 2009.

[3] I. Gijbels, "Censored data," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 2, pp. 178–188, 2010.

[4] R. Hidayat, H. Sumarno, and E. H. Nugrahani, "Survival analysis in modeling the birth

interval of the first child in Indonesia," *Open Journal of Statistics*, vol.4, pp. 198–206, 2014.
[5] Z. Shayan, S. M. T. Ayatollahi, N. Zare, and F. Moradi, "Prognostic factors of first birth interval using the parametric survival models," *Iranian Journal of Reproductive Medicine*, vol. 12, no. 2, pp. 125–130, 2014.
[6] A. Faruk, "Aplikasi regresi Cox pada selang kelahiran anak pertama di provinsi Sumatera Selatan," *Jurnal Matematika*, vol. 7, no. 1, pp. 19–29, 2017.
[7] N. Juhan, N. A. Razak, Y. Z. Zubairi, N. N. Naing, C. H. C. Hussin, and M. A. Daud, "Comparison of stratified Weibull model and Weibull accelerated failure time (aft) model in the analysis of cervical cancer survival," *Jurnal Teknologi (Sciences & Engineering)*, vol. 78, pp. 21–26, 2016.
[8] A. Faruk, A. Amran, and N. Nasir, "Aplikasi model proportional hazard cox pada waktu tunggu kerja lulusan jurusan matematika fakultas mipa universitas sriwijaya," *Jurnal Penelitian Sains*, vol. 17, no. 1, pp. 5–8, 2014.
[9] E. T. Lee and J. W. Wang, Statistical methods for survival data analysis, third edit., John Wiley & Sons, Inc., 2003.
[10] D. G. Kleinbaum and M. Klein, Survival analysis a self-learning text, second edi., Springer, 2005.
[11] A. Gebeyehu, Survival analysis of time-to-first birth after marriage among women in Ethiopia : application of parametric shared frailty model., 2015.
[12] A. Faruk, "Estimasi parameter data tersensor tipe I berdistribusi log-logistik menggunakan maximum likelihood estimate dan iterasi Newton-Rhapson," *Prosiding Seminar Nasional MIPA 2014*, pp. 21–25, 2014.
[13] G. Broström, Event history analysis with R., CRC Press, 2012.