Parametric Survival Model on IPB University's Graduation Data

Muhammad Abdurrasyid Nashiruddin[1], Hadi Sumarno[2], Retno Budiarti[3]
[1,2,3]Department of Mathematics, IPB University, Indonesia
nashiruddin248@gmail.com[1], hadisumarno@apps.ipb.ac.id[2], retnobu@apps.ipb.ac.id[3]

ABSTRACT

Graduation is one of the assessment criteria in the college accreditation process. Students who graduate on time will assist in the assessment of college accreditation. This study aims to determine the distribution that best fits student graduation data and determine the best model to analyze the factors that determine student graduation from IPB University. This study presents some parametric models in survival analysis, specifically, the accelerated failure time (AFT) model and the proportional hazard (PH) model. The objective of this research is to compare the performance of PH model and the AFT models in analyzing the significant factors affecting the student graduation at the IPB University. Based on the study's results, the distribution according to student graduation data is the Burr XII distribution, and the best model using the AIC criteria is the PH Burr XII model. The factors that influence the graduation of IPB University students are gender, faculty, GPA, regional origin, and school status.

—————————— ◆ ——————————

A. INTRODUCTION

Law of the Republic of Indonesia, Number 12 of 2012 concerning Higher Education article, states that a tertiary institution is an academic unit that organizes higher education. The goal of higher education is to produce graduates who are proficient in the fields of science and technology to serve the national interest and enhance the country's competitiveness (Pemerintah Indonesia, 2012). Graduation is one of the assessment criteria in the higher education accreditation process (BAN-PT, 2019). In the undergraduate program, students are expected to graduate within four years (Mendikbud, 2020). However, some students do not graduate on time, so it is necessary to know the distribution of student graduation and the factors that influence it. Therefore, there needs to be a method to explain the time of student graduation. The method can use to explain the time of student graduation is survival analysis (Acton, 2015).

Survival analysis was first developed by the English astronomer Edmund Halley (1656-1742) by describing the longevity of Breslau residents in the 17th century (Kovacheva, 2017). This analysis is essential in demographic, actuarial, and other sciences (Allison, 2014). For example, demographers use this analysis to analyze mortality rates Zoeller et al. (2021),

actuary use this analysis to calculate insurance coverage premiums Staudt & Wagner (2022), and medical professionals use this analysis to estimate the length of life of a patient (Lu et al., 2021). Theories that have discussed resilience analysis include the Kaplan-Meier and Nelson-Altschuler estimation models. Kaplan-Meier and Nelson- Altschuler estimation models are included in the non-parametric study (Moore, 2016). The non-parametric research is relatively easy to understand and apply, but it is less efficient than the parametric analysis when the survival time follows a specific distribution (Lee & Wang, 2013).

Some examples of research that discusses the survival analysis in education is survival models for predicting student dropout at university across time Masci et al. (2022) and survival analysis and its application in education (Syarifuddin, 2012). The model used in this study is still classified as a non-parametric model. Therefore, researchers used a parametric model approach to explain students' study length at the IPB University. The parametric models assume survival times follow a known distribution. Methods of parametric model are such as Tobit, Buckley-James, Penalized regression and Accelerated Failure Time (Atlam et al., 2021).

Parametric regression models focus on the relationship between hazard rate or survival random variables and explanatory variables. There are two common methods for parametric regression model of survival data: proportional hazards (PH) model and accelerated failure time (AFT) model (Wang et al., 2018). The purpose of this research is to determine the best parametric model between PH and AFT model for graduation data IPB University students, as well as to analyze the factors that determine the graduation of IPB University students.

B. METHODS

The following Research Flowchart, as shown in Figure 1.
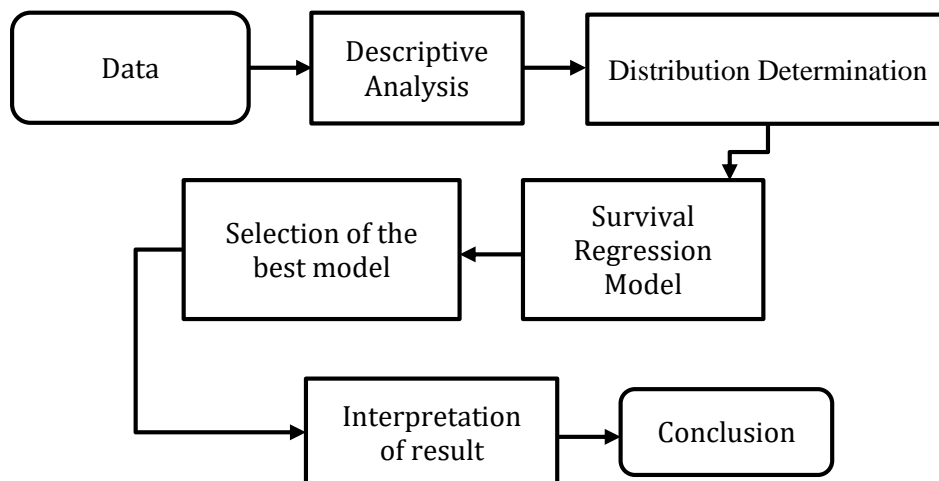
Figure 1. Research Flowchart

1. Data

This stage is carried out a recap of the entire data. The data used is the graduation time data for undergraduate students at IPB University from 2011 to 2016. The data was obtained from the Directorate of Education Administration and New Student Admissions (Dit AP PMB) IPB University.

2. Descriptive Analysis

This stage is carried out descriptive statistics of data to obtain characteristics from the data used. Descriptive statistics consist of numerical variables and categorical variables.

3. Distribution Determination

This stage is to determine the distribution that matches and corresponds to the data using criteria smallest sum squared error (SSE) and Bayesian Information Criterion (BIC). SSE and BIC are among the standard criteria for measuring the model and selecting a good model from a set of potential models. The lower value of SSE and BIC indicates the better the model goodness-of-fit (Pham, 2019). The distribution was limited to five distribution which were Burr, Generalized Extreme Value (GEV), Burr XII, log-logistic, and log-normal distribution. The Burr distribution function is defined as follows:

$$S(x) = \frac{1}{(x^c + 1)^k}, \qquad x \geq 0$$

where $c > 0$ and $k > 0$ are distribution parameters (Almuhayfith et al., 2022).

The Generalized Extreme Value (GEV) cumulative distribution function is defined as follows:

$$F(x) = \begin{cases} \exp\left(-(1 + \theta z)^{-\frac{1}{\theta}}\right), & \theta \neq 0 \\ \exp(-\exp(-z)), & \theta = 0 \end{cases}$$

where $z = (x - \mu)/\sigma$, μ is location parameter with $-\infty < \mu < \infty$, σ is scale parameter with $0 < \sigma < \infty$, and θ is a shape parameter with $-\infty < \theta < \infty$ (Provost et al., 2018).

The Burr XII cumulative distribution function is defined as follows:

$$F(x) = 1 - \left(1 + \left(\frac{x}{a}\right)^c\right)^{-k}, \qquad x \geq 0$$

where $c > 0$, $k > 0$, and $a > 0$ are distribution parameters (Guerra et al., 2021).

The log-logistic survival function is defined as follows:

$$F(x) = \frac{(x/\alpha)^\beta}{1 + (x/\alpha)^\beta}, \qquad x > 0$$

where $\alpha > 0$ and $\beta > 0$ are distribution parameters (Zheng et al., 2021).

The log-normal probability density function is defined as follows:

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(\frac{-(\log x - \mu)}{2\sigma^2}\right), \qquad x > 0$$

where μ and σ are distribution parameters (Awodutire et al., 2022).

4. Survival Regression Model

This stage is to perform a regression analysis of the distribution that matches and corresponds to the data that used. Parametric regression model of survival data is proportional hazards (PH) model and accelerated failure time (AFT) model. Proportional hazards (PH) models can be formulated with or without the assumption of a survival probability distribution (Khan & Khosa, 2015). The hazard function is defined as follows:

$$h(t, \psi) = h_0(t) \exp(\psi) \tag{1}$$

where $h_0(t)$ is baseline hazard function and $\psi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

The accelerated failure time (AFT) model describes the relationship between the response variable and the survival time. In this model, the survival time logarithms are considered a response variable and include an error term assumed to follow a specific probability distribution(Saikia & Barman, 2017). AFT model can be expressed as:

$$S(t|\psi) = S_0(\psi t), \tag{2}$$

where $\psi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and $S_0(t)$ is the baseline survival function.

5. Selection of the best model

This stage is to select and determine the best model between AFT model and PH models by looking at the smallest AIC values. The lower value of AIC indicates the better the model goodness-of-fit (Pham, 2019).

6. Interpretation of result

This stage is Interpret the selected model then analyze the factors which affects graduation. Then compare the results with empirical models. The empirical model used is the Kaplan-Meier model. The Kaplan-Meier (KM) model is a non-parametric analysis to estimate the survival function from a set of survival data. The Kaplan-Meier method estimates conditional survival probabilities computed at specific times determined by events (D'Arrigo et al., 2021). The following equation gives the estimation of the survival function of the model:

$$\hat{S}(t) = \prod_{t_i \leq t}(1 - \hat{q}_i) = \prod_{t_i \leq t}\left(1 - \frac{d_i}{n_i}\right), \tag{3}$$

where n_i and d_i respectively represent the number of subjects at risk at time t_i and the number of individuals who died at time t_i.

7. Conclusion

This stage is to get conclusions and suggestions for the next research.

C. RESULT AND DISCUSSION
1. Data

The data used is the graduation time data for undergraduate students at IPB University from 2011 to 2016. The data was obtained from the Directorate of Education Administration and New Student Admissions (Dit AP PMB) IPB University. The information comprised the length of study at IPB University, gender, origin, faculty, grade point average (GPA), school status, and school origin. Analysis using r and python software.

2. Descriptive Analysis

This section presents the results of descriptive statistics from student graduation survival data. There are 6890 student graduation data used in this study. The descriptive analysis of the categorical data used is presented in Table 1.

Table 1. Descriptive analysis of the categorical variables

Variables	Category	Numbers (%)
Gender (x_1)	Man	38
	Woman	62
Faculty (x_2)	A (x_{21})	11
	B (x_{22})	6
	C (x_{23})	14
	D (x_{24})	6
	E (x_{25})	12
	F (x_{26})	8
	G (x_{27})	20
	H (x_{28})	13
	I (x_{29})	10
Origin (x_4)	Java	75
	Non java	25
School status (x_5)	State	85
	Private	15
School origin (x_6)	MA (x_{61})	6
	SMA (x_{62})	93
	SMK (x_{63})	1

Table 1 explains that the most gender is female and the most faculties are G faculties. The most origin are from Java, the most school status comes from state schools, and the most school origin are from SMA. The characteristics of the sample used represent the student population at the IPB University, as shown in Table 2.

Table 2. Descriptive analysis of the numerical variables

Variables	Min	Mean	Median	Max
Graduation time	3,398	4,361	4,128	7,939
GPA (x_3)	2,250	3,244	3,280	4,000

Table 2 shows the average graduation time is 4,361 years and the average student GPA is 3,244 units. It represents the student population at the IPB University.

3. Distribution Determination

The next step is determining the best distribution of student graduation survival data. Graduation time can be made a pass frequency histogram, as shown in Figure 2.
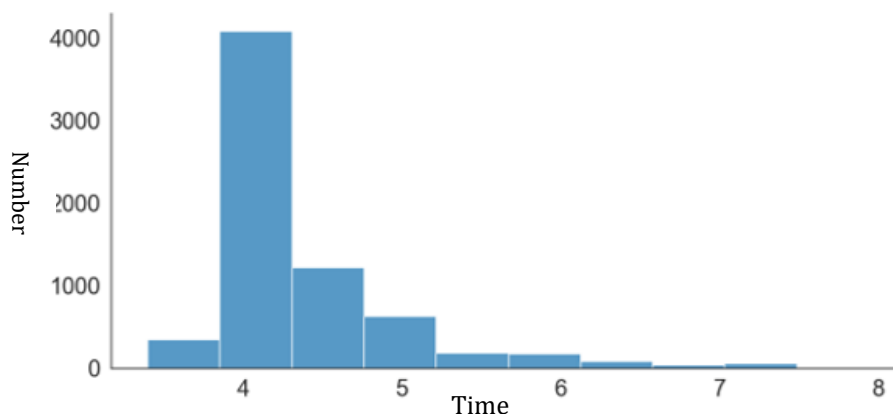
Figure 2. Distribution of data on student graduation time

Figure 2 explains that the highest number of graduates occurs in 4 years because most undergraduate students graduate then. In addition, some are less than four years later, and some are more than four years. Then Figure 1 shows that the distribution is in the form of right skewness so that several distributions have these characteristics, such as the Burr, Burr XII, GEV, loglogistik, and lognormal distributions. Furthermore, the estimation of survival data using this distribution is presented in Figure 3.
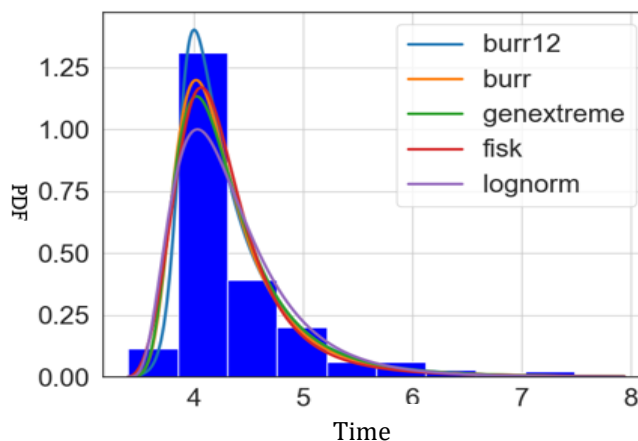
Figure 3. Graph the results of the estimation of several distributions

Figure 3 explains that the most suitable distribution is the Burr XII distribution. This is because the Burr XII distribution has reached the peak of the histogram compared to other distributions. This is reinforced by the results of SSE and BIC for each distribution which are presented in Table 3.

Table 3. AIC Values for Each Model

Distribution	SSE	BIC
Burr XII	0,016273	-89232,05
Burr	0,033285	-84301,42
GEV	0,051813	-81261,30
Fisk (Loglogistik)	0,058059	-80477,08

Lognorm	0,145303	-74156,49

Table 3 explains that the Burr XII distribution produces the smallest AIC and BIC values. Smallest SSE and BIC criteria are used to obtain the most appropriate distribution, so the selected distribution is the Burr XII distribution. Then it will be analyzed using the Burr XII regression model.

4. Survival Burr XII Regression Model

Let X be a random variable that has a Burr XII (c, k, α) distribution with the following survival function:

$$S(x) = \left(1 + \left(\frac{x}{\alpha}\right)^c\right)^{-k},$$

where $c > 0, k > 0$ shape parameter and $\alpha > 0$ is scale parameter (Thupeng, 2016).

a. Accelerated Failure Time Burr XII Model

Suppose given survival time T and an explanatory variable vector X with parameters according to the Accelerated Failure Time (AFT) model can be written as:

$$Y = \beta_0 + \boldsymbol{\beta}'X + \sigma\epsilon \tag{4}$$

then Y is expressed as $log(T)$ so that equation (4) becomes:

$$ln(T) = \beta_0 + \boldsymbol{\beta}'X + \sigma\epsilon \tag{5}$$

based on equation (5), then obtained

$$T = \exp(\beta_0 + \boldsymbol{\beta}'X + \sigma\epsilon) = \exp(\beta_0)\exp(\boldsymbol{\beta}'X) + \exp(\sigma\epsilon)$$

the AFT model of the Burr XII distribution has the survival function as follows:

$$S_{T|X}(t|x) = \left(1 + \left(\frac{t}{e^{\beta_0 + \boldsymbol{\beta}'x}}\right)^c\right)^{-k}. \tag{6}$$

b. Proportional Hazard Burr XII model

The proportional hazard (PH) model is a parametric model of the Cox proportional hazards model. Therefore, the model has the same form as the Cox proportional hazards model. The hazard function at time T, which is affected by the explanatory variable p, is given as follows:

$$h_{T|X}(t|x) = h_0(t)\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) = h_0(t)\exp(\beta_0 + \boldsymbol{\beta}'x)$$

The difference between models is that the baseline hazard function $h_0(t)$ of the PH model is assumed to follow a specific distribution (Muse et al., 2022). In this case, the distribution used is the Burr XII distribution so that the baseline hazard function follows the Burr XII distribution. Therefore, the model can be written as follows:

$$h_{T|X}(t|x) = \frac{\frac{ck}{\alpha}\left(\frac{t}{\alpha}\right)^{c-1}}{\left(1 + \left(\frac{t}{\alpha}\right)^c\right)}\exp(\beta_0 + \boldsymbol{\beta}'x), \tag{7}$$

where $x > 0, c > 0, k > 0$, dan $\alpha > 0$. From equation (7), the survival function of the Burr XII distribution can be written as follows:

$$S_{T|X}(t|x) = \left(1 + \left(\frac{t}{\alpha}\right)^c\right)^{-k\exp(\beta_0 + \boldsymbol{\beta}'x)} \tag{8}$$

5. Selection of the best model

Comparison of AFT model and PH distribution of Burr XII is presented in Table 4.

Table 4. Comparison of the AFT and PH models of the Burr XII distribution

Comparison	AFT Burr XII Model	PH Burr XII Model
Partial likelihood ratio (p-value)	1178,188 (8,47e-243)	1547,911 (2,272e-322)
AIC	6177,027	5809,303

Table 4 explains that the PH Burr XII model and AFT Burr XII model has p-value less than 0,05, so the model is feasible to use. The PH Burr XII model has AIC value 5809,303 and AFT Burr XII model has AIC value 6177,027. So the best model is the PH Burr XII model because it has the smallest AIC value than the AFT model.

6. Interpretation of result

Estimation with the PH Burr XII model to analyze the factors influencing student graduation. Based on Table 4, the partial likelihood test value is 1547,911 with a p-value less than 0,05, so the model is feasible to use. Furthermore, to find out the influencing factors are presented in Table 5.

Table 5. Parameter estimators and p-values of the PH Burr XII model

| Parameter	Estimators	p-value
e^{β_0}	0,35901	-
α	3,90947	-
c	56,58593	-
k	0,16119	-
x_1	0,09706	2,388e-12*
x_{21}	-0,28329	4,931e-21*
x_{22}	0,28630	2,174e-28*
x_{23}	-0,26834	5,897e-21*
x_{24}	0,00510	8,614e-01
x_{25}	0,01450	5,605e-01
x_{26}	-0,08219	2,919e-03*
x_{27}	0,03389	1,156e-01
x_{28}	0,00627	7,862e-01
x_3	0,50795	0,000e+00*
x_4	0,05898	1,554e-05*
x_5	-0,04613	1,341e-02*
x_{62}	0,00126	9,633e-01
x_{63}	0,03565	6,186e-01

*Influence on a significant level of 0,05 and $\chi2$ table with a degree of freedom 1 is 3,841

Table 5 explains that the explanatory variables that have a significant effect on the length of graduation of IPB University students are gender (x_1), faculty (x_2), GPA (x_3), origin (x_4), and school status (x_5). Based on equation (8), the survival function of the PH Burr XII model when graduating students from the IPB University can be written as follows:

$$S_{T|X}(t|x) = \left(1 + \left(\frac{t}{3,90947}\right)^{56,58593}\right)^{-0,16119\,(0,35901\exp(0,09706x_1+\cdots-0,04613x_5))}.$$

For interpretation, because the coefficient of the explanatory variable is inversely proportional to the survival function, if the coefficient is positive, students will pass more quickly. For example, gender (x_1) has a coefficient of 0,09706, meaning that every female student will graduate faster than male students if other explanatory variables are considered the same, as shown in Figure 4.
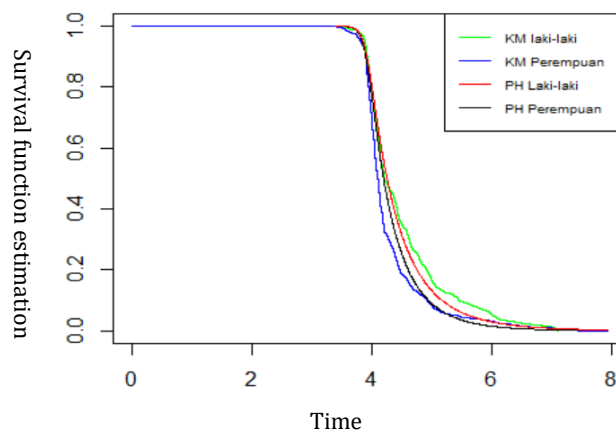
Figure 4. Plot of the estimated survival function of the PH model and the KM model

Figure 4 describes the pattern of the survival function of the KM model and the PH Burr XII model for almost the same. The Kaplan-Meier model seems to be more in line with student graduation data, but this model can only display the survival function per explanatory variable. In contrast to the PH model, which can combine all explanatory variables.

D. CONCLUSION AND SUGGESTIONS

The AFT model and the PH model can be used to determine the factors that influence the graduation time of IPB University students. This study uses the parametric AFT and PH models with the distribution that best fits students' graduation time, the Burr XII distribution. The best model based on AIC criteria is the Burr XII distribution PH model, which produces an AIC value of 5809,303. Analysis of the factors that influence the time of graduation shows that the factors that have a significant effect are gender, faculty, GPA, origin, and school status. Future studies are suggested to use more explanatory variables such as scholarships, pathways of entry, and parents' income.

REFERENCES

Acton, R. (2015). Characteristics of STEM Success: A Survival Analysis Model of Factors Influencing Time to Graduation Among Undergraduate STEM Majors. *Business and Economics Honors Papers*, *1*(1). http://digitalcommons.ursinus.edu/bus_econ_hon/1

Allison, P. (2014). *Event History and Survival Analysis (2nd Edition)*. SAGE Publications, Inc.

Almuhayfith, F. E., Darwish, J. A., Alharbi, R., & Marin, M. (2022). Burr XII Distribution for Disease Data Analysis in the Presence of a Partially Observed Failure Mode. *Symmetry 2022*, *14*(1298), 1–17. https://doi.org/10.3390/sym14071298

Atlam, M., Torkey, H., El-Fishawy, N., & Salem, H. (2021). Coronavirus disease 2019 (COVID-19): survival analysis using deep learning and Cox regression model. *Pattern Analysis and Applications*, *24*(3), 993–1005. https://doi.org/10.1007/s10044-021-00958-0

Awodutire, P. O., Sasanya, B., Ufuoma, O. G., & Balogun, O. S. (2022). Parametric Modelling of Rainfall Return Periods in South Western Nigeria: Survival Analysis Approach. *F1000Research 2022*, *11*(83), 1–14. https://doi.org/10.2139/ssrn.3936312

BAN-PT. (2019). *Akreditasi Perguruan Tinggi Kriteria dan Prosedur 3.0*.

D'Arrigo, G., Leonardis, D., Abd Elhafeez, S., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to Analyse Time-to-Event Data: The Kaplan-Meier Survival Curve. *Oxidative Medicine and Cellular Longevity*, *2021*(2290120), 1–7. https://doi.org/10.1155/2021/2290120

Guerra, R. R., Peña-Ramírez, F. A., & Cordeiro, G. M. (2021). The weibull burr xii distribution in lifetime and income analysis. *Anais Da Academia Brasileira de Ciencias*, *93*(3), 1–28. https://doi.org/10.1590/0001-3765202120190961

Khan, S. A., & Khosa, S. K. (2015). Generalized log-logistic proportional hazard model with applications in survival analysis. *Journal of Statistical Distributions and Applications*, *3*(1), 1–18. https://doi.org/10.1186/s40488-016-0054-z

Kovacheva, T. P. (2017). Life tables - key parameters and relationships between them. *International Mathematical Forum*, *12*(10), 469–479. https://doi.org/10.12988/imf.2017.7225

Lee, E. T., & Wang, J. W. (2013). *Stitistical Methods for Survival Data Analysis* (4 edition). John Wiley & Sons, Inc.

Lu, W., Yu, S., Liu, H., Suo, L., Tang, K., Hu, J., Shi, Y., & Hu, K. (2021). Survival analysis and risk factors in COVID-19 patients. *Disaster Medicine and Public Health Preparedness*, *16*(1), 1–15. https://doi.org/10.1017/dmp.2021.82

Masci, C., Giovio, M., & Mussida, P. (2022). Survival Models for Predicting Student Dropout At University Across Time. *Education and New Developments 2022*, 203–208. https://doi.org/10.36315/2022v1end043

Mendikbud. (2020). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 tentang Standar Nasional Pendidikan Tinggi*.

Moore, D. F. (2016). *Applied Survival Analysis Using R*. Springer International Publishing.

Muse, A. H., Ngesa, O., Mwalili, S., Alshanbari, H. M., & El-Bagoury, A. A. H. (2022). A Flexible Bayesian Parametric Proportional Hazard Model: Simulation and Applications to Right-Censored Healthcare Data. *Journal of Healthcare Engineering*, *2022*(2051642), 1–28. https://doi.org/10.1155/2022/2051642

Pemerintah Indonesia. (2012). *Undang-undang Republik Indonesia Nomor 12 Tahun 2012 tentang Pendidikan Tinggi*.

Pham, H. (2019). A new criterion for model selection. *Mathematics 2019*, *7*(1215), 1–12. https://doi.org/10.3390/MATH7121215

Provost, S. B., Mansoor, M., Saboor, A., & Cordeiro, G. M. (2018). On The q -Generalized Extreme Value Distribution. *REVSTAT-Statistical Journal*, *16*(1), 45–70.

Saikia, R., & Barman, M. P. (2017). A Review on Accelerated Failure Time Models. *International Journal of Statistics and Systems*, *12*(2), 311–322. http://www.ripublication.com

Staudt, Y., & Wagner, J. (2022). Factors Driving Duration to Cross-Selling in Non-Life Insurance: New Empirical Evidence from Switzerland. *Risks*, *10*(1), 187–207. https://doi.org/10.3390/risks10100187

Syarifuddin, A. (2012). *Analisis Survival dan Aplikasinya dalam Bidang Pendidikan (Studi Kasus di Jakarta Selatan)*. Institut Pertanian Bogor.

Thupeng, W. M. (2016). Use of the Three-parameter Burr XII Distribution for Modelling Ambient Daily Maximum Nitrogen Dioxide Concentrations in the Gaborone Fire Brigade. *American Academic Scientific Research Journal*, *26*(2), 18–32. http://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/1273

Wang, X., Yue, Y. R., & Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. CRC Press.

Zheng, X., Chiang, J. Y., Tsai, T. R., & Wang, S. (2021). Estimating the failure rate of the log-logistic distribution by smooth adaptive and bias-correction methods. *Computers and Industrial Engineering*, *156*(1), 107188. https://doi.org/10.1016/j.cie.2021.107188

Zoeller, G. E., Drew, B. L., Schmidt, C. W., Peterson, R., & Wilson, J. J. (2021). A paleodemographic assessment of mortality and fertility rates during the second demographic transition in rural central Indiana. *American Journal of Human Biology*, *34*(1), 1–21. https://doi.org/10.1002/ajhb.23571