Volume 6 No 2 Desember (2023)

e-ISSN: **2615-3270**

p-ISSN: **2615-3599**

Eigen Mathematics Journal

Homepage jurnal: http://eigen.unram.ac.id

Survival Analysis with Cox Proportional Hazard Regression Modeling (Case Study: Student Study Period Data on Engineering Faculty of Bangka Belitung University)

Ineu Sulistiana[a,*], Elyas Kustiawan[b], Ririn Amelia[c]

[a]Universitas Bangka Belitung, Jalan Kampus Terpadu Universitas Bangka Belitung, Kelurahan Balunijuk, Kecamatan Merawang, Kabupaten Bangka 33172, Indonesia. Email : ineu.sastrawinangun90@gmail.com
[b]Universitas Bangka Belitung, Jalan Kampus Terpadu Universitas Bangka Belitung, Kelurahan Balunijuk, Kecamatan Merawang, Kabupaten Bangka 33172, Indonesia.: elyaskustiawan@gmail.com
[c]Universitas Bangka Belitung, Jalan Kampus Terpadu Universitas Bangka Belitung, Kelurahan Balunijuk, Kecamatan Merawang, Kabupaten Bangka 33172, Indonesia.: ririn-amelia@ubb.ac.id

A B S T R A C T

Student study time is the time needed by students to complete their education, which starts from the time they enter college until they are declared graduated or have completed their study period. In the study period data, survival time observations were only carried out partially or not until the failure event. In other words, termination occurs until the observation deadline. This termination occurred due to several factors that allegedly influenced the student's study period. This study intends to determine what variables influence the study period of students of the Faculty of Engineering, University of Bangka Belitung through survival analysis. Using study period data for students of the Faculty of Engineering, University of Bangka Belitung, class of 2015/2016, this study used the Kaplan Meier Estimation to see the survival function of each factor causing the length of the study period graphically and the Log Rank Test statistically. Meanwhile, to look at the factors that determine the length of a student's study period, researchers used the Cox Regression and Maximum Likelihood Estimation (MLE) models to find the best model. The results of the data analysis show that there are differences in the survival function in each category for all variables graphically, while the statistical comparison of the results of the estimation of the survival function curve based on gender and organizational status is not significantly different. The results of the analysis also show that the proportional hazard assumption is fulfilled through the cumulative hazard log so that categorical variables can be used in the Cox Regression model. Based on the results of the likelihood estimation, the variables that have a significant effect on the study period of Engineering Faculty students are majors and GPA variables. Furthermore, from the interpretation of the model parameters, it is obtained that the Hazard Ratio (HR) value for the study period of Mechanical, Mining and Electrical Engineering students is faster than that of Civil Engineering students, while students with GPA ≥ 3.00 have a shorter study period than students with GPA < 3.00.

* Corresponding author.
Alamat e-mail: ineu.sastrawinangun90@gmail.com

Keywords: Kaplan Meier Estimation, Log Rank Test, Cox Regression, Maximum Likelihood Estimation, Hazard Ratio, Study Period
Diserahkan: 02-05-2023; Diterima: 10-08-2023; Doi: https://doi.org/10.29303/emj.v6i2.170

1. Introduction

Censored data is observational data from the object under study at a certain time and does not fail until the research ends (Kusumawardhani, et al., 2018). Censorship in an observation will result in incomplete information on the length of time or duration of the data obtained. One of the causes of censored data is termination, which occurs when the research period ends while the observed object has not yet reached the failure event (Pyke, D and Thompson, 1986). In the study period data, survival time observation was only carried out partially or not until the failure event. In other words, termination occurs until the observation deadline. This termination occurred because many factors were thought to influence the student's study period, both internal and external factors. Internal factors are factors that come from within the student such as learning ability, level of student activity, ability to solve problems (level of intelligence), and others. While external factors are factors that come from outside the student's self, such as environmental conditions, association, the amount of parental support, infrastructure and facilities owned, and others (Fitriana, 2016; Fitriani, 2018). This study intends to determine what variables influence the study period of students of the Faculty of Engineering, University of Bangka Belitung through survival analysis. Survival analysis is a statistical method that can be used to answer the question of whether and when an interesting event occurs. (Guo, 2010). In addition, survival analysis is used to analyze data that aims to find out the results of the variables that influence an event from the beginning to the end of the event (Kleinbaum & Klein, 2011). Based on the definition and characteristics of survival analysis, there are several examples of data that can be used as survival data. One of them is data on the student's study period, where the study period is the time needed by students to complete their education starting from the time they enter college until they are declared graduated or have completed their study period. In this study, the relationship between the dependent variable, namely the endurance of students at the Faculty of Engineering, University of Bangka Belitung, and the explanatory variables, namely gender, major, GPA, scholarship status, participation in organizations, and sensory status using the Cox Regression model approach and the Maximum Likelihood Estimation (MLE) method in selecting the best model. In addition, the estimation of the survival function graphically is carried out through Kaplan Meier estimation, then followed by the Log Rank test to test whether or not there is a difference in the Kaplan Meier survival curve in variables that have two or more categories (Kleinbaum and Klein, 2005).

2. Methodology

2.1. Research environment space

This research is only focused on data student from the Faculty of Engineering, University of Bangka Belitung who registered at the beginning of the 2015/2016 academic year odd semester (2015.1) to the 2019/2020 academic year even semester (2019.2). Furthermore, to determine the relationship between the survival time of students at the Faculty of Engineering, University of Bangka Belitung, and its explanatory variables, this study uses the Cox Regression model approach. Where the Cox Regression model is one of the methods in survival analysis that links responses in the form of survival time with explanatory variables. The following are the variables used in the research:

1. Endurance time of student study period. This variable is the response/ dependent variable which is observed from the time students carry out their studies until they are declared to have passed their Bachelor degree whic is denote by "t" snd the unit of time is the semester. This variable is measured in semester units with the following conditions:
   a. If a student is declared to have passed until the even semester of the 2019/2020 academic year, the survival time is declared as uncensored data.
   b. If the study period exceeds the even semester of the 2019/2020 academic year, it is declared as censored data.
2. Gender (male = 1; female = 0)
3. Department (civil engineering = 1; machine engineering = 2; mining engineering = 3; electrical engineering = 4).
4. Grade Point Average (GPA) (GPA < 3.00 = 0; GPA ≥ 3.00 = 1).
5. Scholarship Status (ever = got a scholarship 1; never = 0).
6. Organizations status (ever = join a

student organizations = 1; never = 0)
7. Censorship status (uncensored data = 1; censored data = 0). The sensor used in this study is the right sensor type 1 (time sensor) because the research time is set at a certain time interval, so that students who do not experience events within that time interval cannot be determined with certainty about the duration of their study period.

2.2. Data analysis

2.2.1. Kaplan meier estimation and log rank test

To estimate survival function S(t) you can use the Kaplan- Meier estimator or often also called the Product-Limit estimator as follows:

$$\hat{S}(t) = \begin{cases} 1 & jika\ t < t_i \\ \prod_{ti \le t}(1 - \frac{d_i}{Y_i}) & jika\ t_i \le t \end{cases} \quad (2.1)$$

Where d_i is the number of events and Y_i is the number at risk. The Kaplan-Meier estimator is a function of the ladder that goes down when there is an event. The Kaplan-Meier estimator is non-parametric in the sense that it does not assume a finite number of parameters. The number of parameters or quantities to be estimated in Kaplan-Meier is as many as the points in time where the event occurs (Andardono, 2012). One method that is often used in nonparametric survival analysis is Kaplan Meier analysis followed by the Log Ranktest. Kaplan Meier analysis is used to estimate thesurvival function. Then from the estimation of the survival function a Kaplan Meier survival curve can be formed. Meanwhile, the Log Rank test is used to test whether there is a difference or not in the Kaplan Meier survival curve for variables that have two or more categories. With the hypothesis for the Log Rank test as follows:

H_0: there is no difference between survival curves

H_1: there is at least one difference between the survival curves

The test statistics used in the Log Rank test are divided into the Log Rank test for two groups andthe Log Rank test for more than two groups. The test statistics for the two-group Log Rank test refers to:

$$\text{Log rank statistic} = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \quad (2.2)$$

The test statistics for the Log Rank test for morethan two groups are as follows:

$$Log\ rank\ statistic = d'V^{-1}d \quad (2.3)$$

or with the approximation formula Log Rankstatistics

$$\mathcal{X}^2 = \sum_i^G \frac{(O_i - E_i)^2}{E_i} \quad (2.4)$$

Hypothesis H_0 will be rejected if the *p-value* lessthan α atau Log rank$_{\text{statistic}} \approx \mathcal{X}^2{}_{\text{count}}$ more than $\mathcal{X}^2{}_{\alpha,df}$ with degrees of freedom equal to G-*1* (Kleinbaum and Klein, 2005; Suhartini, et al.,2018).

2.2.2. Proportional hazard assumption

There are three ways to test the Proportional Hazard (PH) assumption, namely by using a graphical approach using the Log Minus Log (LML) Survival plot, using the Schoenfeld residual and by adding a time dependent variable. In using the Log Minus Log (LML) Survival plot, survival data are grouped according to the level of one or more factors. If the variable is continuous then its value needs to be grouped into categorical variables. The log minus log survival plot is a plot of the logarithm of the estimated cumulative hazard function for survival time, which will produce a parallel curve if the rate of proportional hazard is across different groups (Collett, 2003) (Iskandar, 2015). Furthermore, if the graphs or plots of LML Survival between categories in one independent variable are parallel or do not intersect, then the PH assumption is fulfilled and the categorical independent variables can be used in the Cox PH semiparametric regression model (Kleinbaum and Klein, 2010; Chandra and Rohmaniah, 2019). This study uses a cumulative hazard plot to test the PH assumption.

2.2.3. Regresi cox model and maximum likelihood estimation (MLE)

In determining the relationship between variables, this study uses the Cox Regression model. Where in this model does not require any assumptions or information on the distribution of survival data. The Cox regression model is a proportional hazard regression model with the

baseline hazard function modeled non-parametrically and the independent variable function modeled parametrically. Cox regressionis modeled as follows:

$$h(t|x) = h_0(t)\varphi(X,\beta) \quad (2.5)$$

with $x = (x_1, \dots, x_2)$ is a covariate vector (independent variable) and β'= (β_1, …, β_p) is the parameter of the regression model. In this regression the hazard for each individual is the same as the baseline hazard $h_0(t)$ if the effect of the independent variables is not taken into account, or the values x = (x_1, …, x_2) are all equal to zero. The hazard of each individual is modified multiplicatively by the characteristics of each individual, which is expressed by $\varphi(x,\beta)$ (Andardono, 2012). Furthermore, parameter significance testing is carried out using likelihood estimation. This is done to see whether or not there is an influence of the covariate variables on the dependent variable. With the form of a hypothesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_1: *there is at least one* $\beta_j \neq 0$, j = 1, 2, …, p

If p-value less than 0,05, then H_0 rejected. In other words, all independent variables have a significant effect on the model (Collett, 2015; Chandra and Rohmaniah, 2019). While the selection of the best model is doneusing backward selection. Initially allindependent variables were included in themodel equation, then excluded one by one based on the greatest p-value. Overall if all the p-value of each variable included in the model are significant then the backward selection is stopped (Fitriani, 2018).

2.2.4. Hazard ratio (HR)

The assumption of a constant Hazard Ratio (HR) is the underlying assumption for Cox Regression. In the proportional hazard assumption, all individuals are considered to have the same baseline hazard then the value is different or modified according to the characteristics or information of each individual (Andardono., 2012). In other words, the parameters in the Cox Regression can beinterpreted as a hazard ratio. According to Lee and Wang in 2003, and Fitriani in 2018, the Hazard Ratio (HR) can show an increase or decrease in the risk of individuals subjected to certain treatments. Suppose there are two individuals with certain characteristics, then from the general cox proportional hazard equation, the formula for estimating the hazard ratio is obtained as follows:

$$HR = \frac{h(t|X_j)}{h(t|X_0)} = \frac{h_0(t)\exp(\beta X_1)}{h_0(t)\exp(\beta X_0)} = \frac{\exp(\beta X_1)}{\exp(\beta X_0)}$$

$$= exp^{(X_1 - X_0)\beta}, \forall t > 0 \quad (2.6)$$

There are 3 kinds of provisions regarding the increase or decrease in the hazard value as follows.

1. $\beta_j > 0$ then every increase in value x_j will increase the hazard value or the greater the risk of an individual to experience an event.
2. $\beta_j < 0$ then every increase in value x_j will reduce the hazard value or the smaller the riskof an individual experiencing an event.
3. $\beta_j = 0$ then the risk of an individual to experience the same event with the risk of anindividual to fail.

3. Results and Discussion

3.1. Kaplan Meier estimation and log rank test

The following is the Kaplan Meier survival curve from the study period data of Engineering Faculty students 2015/2016:

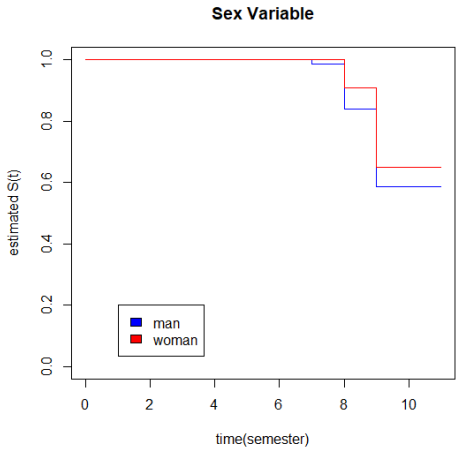

Figure 1. Gender Variable Survival Function Estimation Curve

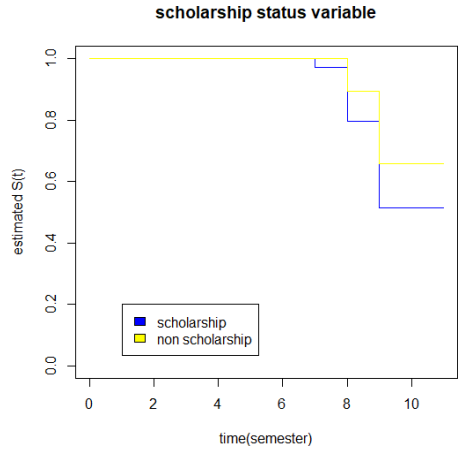

Figure 2. Variable Survival FunctionEstimation Curve of Scholarship Status

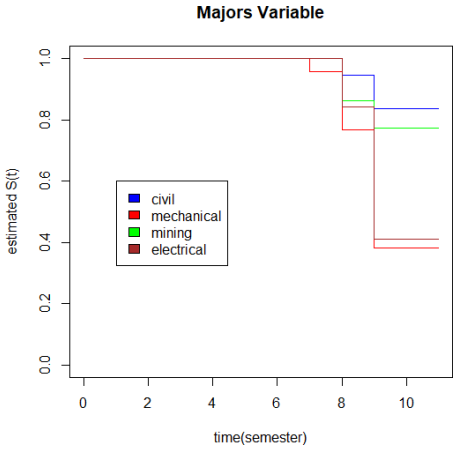

Figure 3. Variable Department Survival Function Estimation Curve

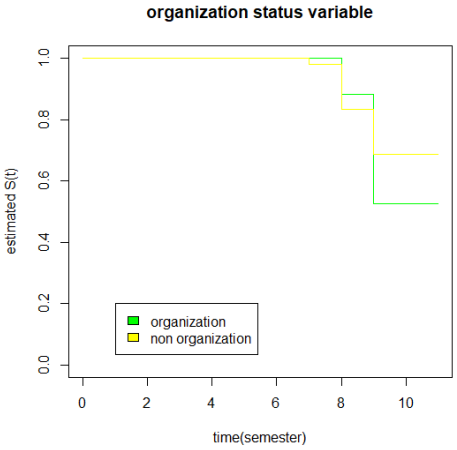

Figure 4. Survival Function Estimation Curve for Organizational Status Variables

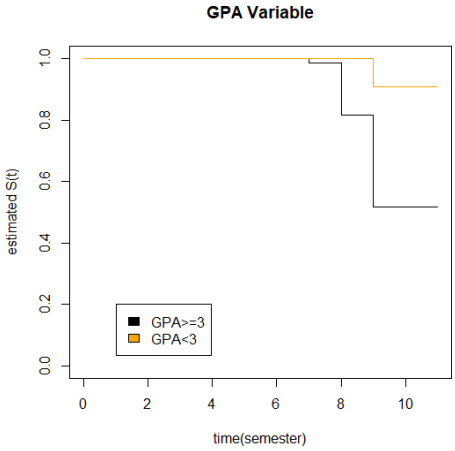

Figure 5. Survival Function Estimation Curve for GPA Variables

Graphical comparison of the survival function of each variable was carried out through the Kaplan Meier curve. The following describes each curve:

Figure 1: The study period for male students is shorter than female students.

Figure 2: The study period of students who have received a scholarship is faster than the status of students who have never received a scholarship.

Figure 3: The study period for Mechanical Engineering students is shorter than other majors. While the study period for Civil Engineering students is longer than other majors

Figure 4: The study period of students who have joined student organizations is shorter than students who have never joined studentorganizations.

Figure 5: The study period of students with GPA ≥ 3 is faster than students with GPA < 3.

From the explanation that has been described, it is suspected that there are differences in the survival function in each category of all variables. Next, a statistical comparison of the survival function will be carried out. This study uses the log rank test to test whether there is a difference or not in the Kaplan Meier survival curve for variables that have two or more categories. Following are the results of the log rank test for each categorical variable:

Table 1. Log Rank Test

Variable	Log Rank	df	p-value	Decision
Gender	0.9	1	0.4	H0 failed to reject
Department	32.7	3	0.0000004	H0 is rejected
Scholarship Status	4.9	1	0.03	H0 is rejected
Organization status	3.4	1	0.06	H0 failed to reject
GPA	21.1	1	0.000004	H0 is rejected

The results of the statistical log rank test on the variables of gender and organizational status obtained a p-value greater than the level of significance (α) 0.05, This means that the comparison of the results of the estimation of the survival function curve based on the variables of gender and organizational status is not significantly different. While the results of the statistical log rank test on major variables, scholarship status and GPA were obtained p-value less than the significance level (α) 0.05, This means that the comparison of the results of the estimation of the survival function curve based on the variables of majors, scholarship status and GPA is significantly different.

3.2. Proportional hazard assumption

The following plots the estimation of the cumulative hazard function for testing the proportional hazard assumption as another form of the Log Minus Log plot (LML) survival.

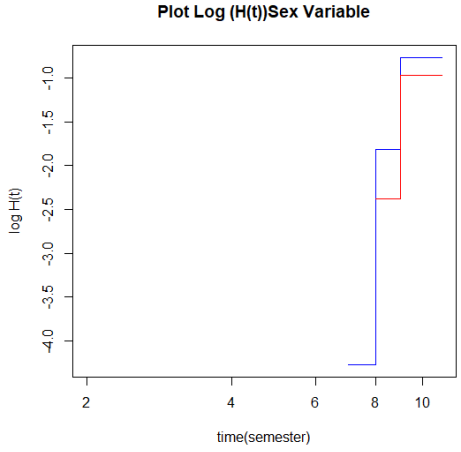

Figure 6. Gender Variable Cumulative Hazard Log Curve

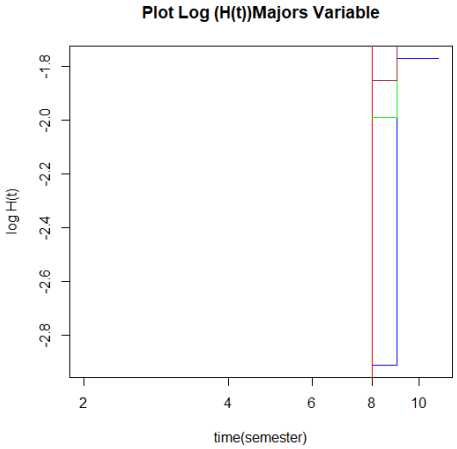

Figure 7. Major Variable Cumulative Hazard Log Curve

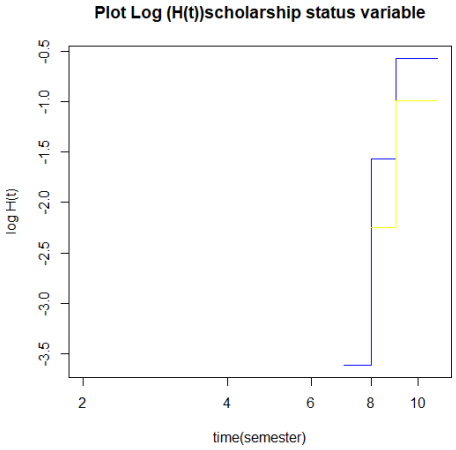

Figure 8. Scholarship Status Variable Cumulative Hazard Log Curve

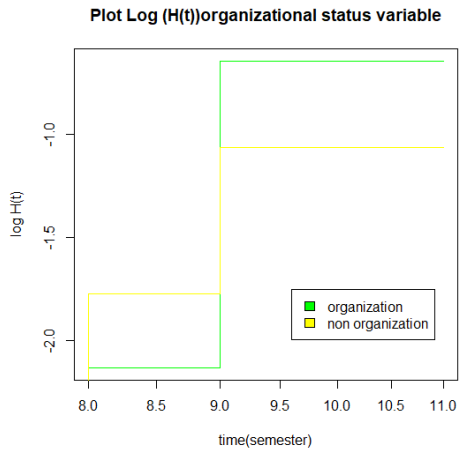

Figure 9. Organizational Status Variable Cumulative Hazard Log Curve

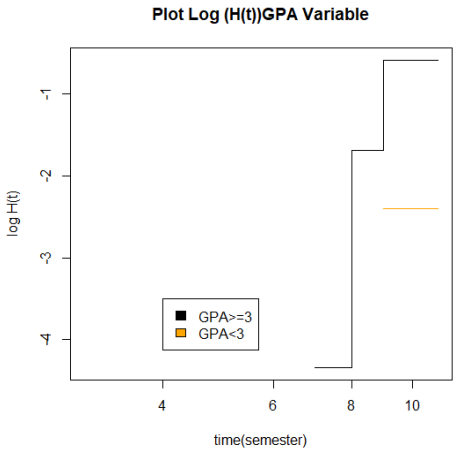

Figure 10. GPA Variable Cumulative Hazard Log Curve

Figure 6, 7, 8, 9 and 10 shows a curve or plot between categories in one independent variable that is parallel or does not intersect, thenthe Proportional Hazard assumption is fulfilled and the categorical independent variables can be used in the Cox Proportional Hazard regression model.

3.3. Cox regression model and maximum likelihood estimation

This study uses the Cox Regression and Maximum Likelihood Estimation model to determine the relationship between the variables of gender, major, scholarship status, organizational status and GPA on student's study period as their survival time. As for selecting the best model, this study uses backward elimination which is one of selecting variables that enter or leave the model (Collett, 2003). The following is the R output for the best model based on the regression results and likelihood estimation:

Table 2. Parameter Estimation Table of the Best Cox Regression Model

Variable	Estimation (β)	Sig.
Mechanical Engineering	1.6238	0.0000216
Mining Engineering	0.0581	0.89960
Electrical Engineering	1.0811	0.00455
GPA ≥ 3.00	2.0638	0.0000780

Source: R Output for Cox Regression

Table 3. Estimation of Likelihood for the Best Model

Variable	Df	AIC	LRT	Sig.
Major	3	788.54	31.889	0.0000005524
GPA	1	787.91	27.260	0.0000001779

Source: Output R for Estimation of Likelihood

Based on the R output for likelihood estimation, with the variables majoring in Civil Engineering and GPA < 3.00 as the reference category, it is obtained that the p-value of the major and GPA variables is less than the significance level (α) 0.05 or in other words the major and GPA variables have a significant effect on the hazard function of the student's study period. So that the majors and GPA variables are feasible to be included in the regression model.

The following is the Cox regression model on study period data for Engineering Faculty students 2015/2016:

$$h_i(t|X) = h_o(t) \exp\,(1{,}6238_{mechanical} + 0{,}0581_{mining} + 1{,}0811_{electrical} + 2{,}0638_{GPA\geq 3.00})$$

3.4 Hazard ratio

Parameters in the Cox Regression can be interpreted as a Hazard Ratio (HR). The

regression model shows the parameters of each major and GPA variable category are positive, or $\beta_j > 0$ so that the greater the risk of an individual experiencing an event (Fitriani, 2018). The following is the Hazard Ratio (HR) value for each category of majors and GPA variables:

Table 4. Hazard Ratio Value

Variable	Estimation (β)	Hazard Ratio (exp (β))
Mechanical Engineering	1.6238	5.072
Mining Engineering	0.0581	1.060
Electrical Engineering	1.0811	2.948
GPA ≥ 3.00	2.0638	7.876

Source: Output R for Hazard Ratio

1. For the Mechanical Engineering Department variable, an HR value of 5.072 is obtained, meaning that the study period of Mechanical Engineering students is 5.072 times fasterthan Civil Engineering students.
2. For the variable majoring in Mining Engineering, an HR value of 1.060 is obtained, meaning that the study period of Mining Engineering students is 1.060 times faster than Civil Engineering students.
3. For the variable majoring in Electrical Engineering, an HR value of 2.948 is obtained, meaning that the study period of Electrical Engineering students is 2.948 times faster than Civil Engineering students.
4. For the GPA variable, the HR value was 7.876, meaning that the study period of students with GPA ≥ 3.00 was 7.876 times faster than students with GPA < 3.00.

4. Conclusion

Based on the results of data analysis and discussion, the Cox Regression model is obtained for the 2015/2016 Faculty of Engineering student data as follows:

$$h_i(t|X) = h_o(t) \exp(1{,}6238_{mechanical} + 0{,}0581_{mining} + 1{,}0811_{electrical} + 2{,}0638_{GPA\geq 3.00})$$

The variables that have a significant effect on the study period of students of the Faculty of Engineering 2015/2016 are majors and GPA variables. Furthermore, based on the Hazard Ratio (HR) value, the study period of students from the Department of Mechanical, Mining and Electrical Engineering is faster than the study period for students from the Department of Civil Engineering. While students withGPA ≥ 3.00 have a shorter study period than students with GPA < 3.00.

Acknowledment

The author expresses his gratitude for the financial support provided by the University of Bangka Belitung through the 2020 Departmental Level Lecturer Research scheme.

REFERENCE

Chandra, N. E., dan Rohmaniah, S. A. (2019). Analisis Survival Model Regresi Semiparamterik pada Lama Studi Mahasiswa. *Jurnal Ilmiah Teknosains*, V, 94 – 98.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall, US.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall, USA.

Danardono. (2012). *Analisis Data Survival*. Diktat Kuliah Program Studi Statistika Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Gajah Mada.

Fitriana, R. (2016). *Analisis Survival Faktor-faktor yang Mempengaruhi Lama Studi Mahasiswa Pendidikan Matematika Angkatan 2010 dengan Metode Regresi Cox Proportional Hazard*. Skripsi FMIPA Universitas Negeri Semarang, Semarang.

Fitriani, I. D. (2018). *Analisis Regresi Cox Proportional Hazard pada Identifikasi Faktor-Faktor yang Mempengaruhi Lama Studi Mahasiswa S1 Fmipa Universitas Islam Indonesia.* Skripsi FMIPA Universitas Islam Indonesia, Yogyakarta.

Iskandar, B. M. (2015). *Model Cox Proportional Hazard pada Kejadian Bersama*. Skripsi Universitas Negeri Yogyakarta, Yogyakarta.

Kleinbaum, D. G., dan Klein, M. (2005). *Survival Analysis: A Self-Learning Text* (2nd ed). Springer, New York.

Kleinbaum, D. G., dan Klein, M. (2010). *Survival Analisis Third Edition*. Springer, New York.

Kusumawardhani, G. E., Suyono dan Santi, V. M. (2018). Analisis Survival dengan Model Regresi pada Data Tersensor Berdistribusi Log-logistik. *Jurnal Statistika dan Aplikasinya (JSA)*, **2,** 28-35.

Lee, E. T., dan Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis Third Edition*. John Wiley & Sons, Inc, New Jersey.

Pyke, D dan Thompson, J. (1986). Statistical Analysis of Survival and Removal Rate Experiments. *Ecological*, 67, 240-245.

Suhartini, A., Rahmawati, R., dan Suparti. (2018). Analisis Kurva Survival Kaplan Meier Menggunakan Uji Log Rank (Studi Kasus: Pasien Penyakit Jantung Koroner di RSUD Undata Palu). *Jurnal Gaussian*, 7**,** 33-42.