ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 10 , October 2017

Survival Analysis for Undergraduate Student’s Performance

Azme Khamis, Che Azmeeza Che Hamat, Mohd Asrul Affendi Abdullah, Sabariah Saharan, Norhaidah Mohd Asrah

Department of Mathematics and Statistics, Faculty of Applied Science and Technology,
Universiti Tun Hussein Onn Malaysia, Johor,Malaysia

ABSTRACT: The goal of this study is to investigate undergraduate students’ performance using survival analysis. Kaplan Meier estimator is used to examine the significant differences among the Grade Point Average (GPA) obtained and identified which parametric survival model gave the best fitted to the data. The entry qualification for student enrols to degree program at UTHM is matriculation, diploma and Malaysian higher certificate education or STPM. The Cumulative Point Average (CPA) student for 2010/2011 intake was considered in this study. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to obtain the best parametric survival model. It was found that Weibull model was outperformed compare to Kaplan Meier estimator, and STPM students had the highest cumulative survival probability and performed significantly better than diploma and matriculation.

KEYWORDS: Survival Analysis, Kaplan Meier, Weibull Model

I. INTRODUCTION

Students’ pre-university academic performance had not being able to predict the academic performances in undergraduate study at university [1]. The students whose work hard, be responsible and had the strong commitment to learning during the study had a good potential that can be polished with a proper guided from the academic staff [2]. A numerous number of studies had carried out in search the factors that affecting students’ performance in undergraduate program. The previous result in pre-university did not influence the students result during undergraduate program [3]. This researcher also stated that the students ‘grade point average (GPA) for every year of study were changing and did not constant. The students’ performance obtained by measuring the learning assessment and curriculum [4]. Generally, most of higher learning institutions in Malaysia used the cumulative grade point average (CGPA) to evaluate students’ performance. The most of the previous studied mentioned about students’ achievement during graduation being the measure of students’ success [5]. The entry qualification for students to pursue undergraduate program at UTHM are diploma, matriculation or Malaysian Higher Education Certificate (STPM).

II. SIGNIFICANT OF THIS SYSTEM

The goal of this study is to investigate undergraduates’ students’ performance using survival analysis. The proposed model based on survival analysis can be used to estimate the students’ survival and performance. The proposed model can be used to help the academic staffs in monitoring students’ achievement and take any earlier action in order to help students to maintain or improved their performance. The statistical packages used for this research were SPSS and R package.

III. LITERITURE REVIEW

The survival analysis primarily applied to the medical field. However, the methods of survival analysis with repeated events to a longitudinal data set to illustrate the studied on survival and hazard function for several groups of college students too [6]. Some common term used in survival analysis was censored, survival function and hazard function. Censoring occurred when a person lost to follow up during the study period and had failed during the study. Survival function gave the probability that a student performed on specified time, *t* and a fundamental to a survival analysis. Hazard function gave the instantaneous potential per unit time for the event occurs given the student had performed to time t. The survival analysis applied in an academic sector and used to measure student achievement and determine whether the categorical variables had different survival time using survival function analysis [7].

Data mining was one of the famous techniques to analyse the students' performance and widely used in an educational area recently [8]. The process is also known as educational data mining, which was a process used to extract useful information and patterns from a huge educational database [9] and would assist the educator in providing an effective approach. The data for students GPA were incomplete since had numbers of students were not finish the study and the students data were lost to follow up until the end of the study. The incomplete data means that the data was censored. Many of the students tend to drop out of the study before study end means the data was the right censored data. There was difficulty to handle the censored data. For this situation, survival analysis allowed inclusions of information for those students who were censored and had incomplete data [10]. The advantage of survival analysis is that to overcome the difficulty of handling censored observations [11]. The aim of this study is to compare the survival and hazard rate of students based on the level of qualification used Kaplan Meier estimator and the parametric survival model.

IV. METHODOLOGY

A. Kaplan-Meier: The Kaplan Meier estimator used to compute and plot survival probability for the Undergraduate students' performance (GPA). Nonparametric approach to estimate the survival function using standard Kaplan Meier (KM) estimator [12] [13].For the parametric survival model, the three distribution function which was Exponential, Weibull and Gamma model will use for this study. The best parametric model will be selected based on the minimum value of the Akaike Information Criterion(AIC) and Bayesian Information Criterion (BIC) and the survival probability plot as a diagnosed tool to assess whether a parametric distribution fitted to the survival data [14].

The Kaplan-Meier (product limit) method was a special case of the life table technique, in which the series of time intervals were formed in such a way that only one event occurred in each time interval [16].Suppose that there are n individuals with observed survival times $t_1, t_2, t_3 \dots, t_n$ and r was the failure time among the individuals where $r < n$. The ordered failure times

$$t_{(j)}, j = 1, 2, \dots, r = t_{(1)} < t_{(2)} <_{\dots} < t_{(r)} \qquad (1)$$

Let n_j which $j = 1,2, \dots, r$ be the number of individuals who were success just before just before the time $t_{(j)}$ and let $d_{(j)}$ be the number of individuals failed at time $t_{(j)}$. The probability for an individual failed during the interval $t_{(j-1)}$ to $t_{(j)}$ was estimated as $\frac{d_{(j)}}{n_{(j)}}$. Therefore the corresponding estimated survival probability in that interval was $\frac{\left(n_{(j)} - d_{(j)}\right)}{n_{(j)}}$ .If the censored survival times and one or more failure times were same, then in this case it was assumed that the censored survival time was taken to occur immediately after the failure time. So, the estimated survival function for any time in the t in the j^{th} constructed time interval from $t_{(j)}$ to $t_{(j+1)}, j = 1, 2, \dots, r$ and all the preceding time intervals was led to the following Kaplan –Meier estimate of the survivor function,

$$\hat{S}(t) = \prod_{t \le t_{(j)}} \frac{n_j - d_j}{n_j} \qquad (2)$$

for $t_{(j)} \le t < t_{(j+1)}, j = 1, 2, \dots, r, \hat{S}(t) = 1$ for $t < t_{(1)}$, $\hat{S}(t) = 0$ for $t \ge t_{(r)}$ if $t_{(r)}$ was the last observation.

It noted that the Kaplan-Meier estimate was unspecified for the largest censored survival time. It was important to note that the Kaplan-Meier estimator in spite of its simplicity and wide applicability.

B. Exponential model: The first parametric model was the exponential model. This model was characterized by a constant hazard rate, λ and it was the only parameter in this model. A large value ofλ indicates high risk and short survival. When the survival time T followed the exponential model with a parameter λ, the probability density function,$f(t)$defined as

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \ge 0, \lambda > 0 \\ 0 & t < 0 \end{cases} \qquad (3)$$

and the survival function, $S(t) = e^{-\lambda t}$ for $t \ge 0$. The hazard function is define as, $h(t) = \frac{f(t)}{S(t)}$. Then it can be simplified as, $h(t) = \lambda,$for $t \ge 0$ where it will be a constant and independent of t.

C.Weibull model: The Weibull model was a generalization of the exponential distribution. The model is characterized by two parameters which are γ and λ where γis the shape and $\lambda$is scale parameters. The probability density function, $f(t)$were respectively

$$f(t) = \lambda\gamma\,(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}}t \geq 0,\ \gamma,\ \lambda > 0 \tag{4}$$

The survival function, $S(t)$is define as$S(t) = e^{-\lambda(t^{\gamma})}$and the hazard function, $h(t) = \lambda\gamma\,(\lambda t)^{\gamma-1}$. The survival curve, $log_e S(t) = -(\lambda t)^{\gamma}$, if$\gamma = 1$, $thelog_e S(t)$ is a straight line. If$\gamma < 1, log_e S(t)$ decreases very slowly from zero and then approaches a constant value,if$\gamma > 1$, $log_e S(t)$ decreases sharply from zero as the value of$t$ increases.

D. Gamma model: This model described survival patterns that had constant initial hazard rates. The gamma model was limited used in survival analysis and did not had closed form expressions for survival and hazard functions. Both included the incomplete gamma integral.

$$I_{\gamma}(x) = \frac{\int_0^x S^{\gamma-1}e^{-s}ds}{\Gamma(\gamma)} \tag{5}$$

Consequently, traditional maximum likelihood estimation was difficult and required the calculation of such incomplete gamma integrals, which imposed additional numerical problems in parameter estimation. A random variable,X was gamma distributed with shape parameter, γ and scale parameter, λ.

$$f(t) = e\frac{\lambda^{\gamma}x^{\gamma-1}e^{-\lambda x}}{\Gamma(\gamma)}, \text{where } \gamma, \lambda > 0 \tag{6}$$

If $\gamma = 1$, the gamma distribution was reduced to the exponential distribution.

Then the survival function, $S(t) = 1 - I_{\gamma}(\lambda x)$ and the hazard function,$h(t) = \frac{\lambda^{\gamma}x^{\gamma-1}e^{-\lambda x}}{(1-I_{\gamma}(\lambda x))\Gamma(\gamma)}$.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to measure the model performance. The AIC defined as, $AIC = -2 * \log(likelihood) + 2(k)$,where $k$is the number of parameters included in the model, k= 1 for the exponential model, k = 2 for the Weibull and Gamma model. The model with smallest value was termed as the preferred and best fit model [17].

Bayesian Information Criteria (BIC) is defined as, $BIC = -2 * \log(likelihood) + qln(N)$, where q is the number of covariates in the model, where q = 1 for the exponential model, q = 2 for the Weibull and Gamma and N was the number of observations in the data set. The model with the smallest values was chosen as the best model.

V. RESULT AND DISCUSSION

A.Kaplan Meier Analysis: The Kaplan Meier result summarizes survival data in terms of the number of events and the proportion surviving at each event time point. The time referred to the number of semesters enrolled by undergraduate students. The event from this study was students' achievement GPA > 2.0 during degree based on a number of the semester enrolled by students. Table 1 presents the result for eight semesters of students' achievement study based on qualification.

Table 1: Kaplan Meier comparison result for eight semesters of students' achievement study based on qualification entry

Semester enrolled	Diploma			Matriculation			STPM		
	No. of risk	No of event	Cum. Survival probability	No. of risk	No of event	Cum. Survival probability	No. of risk	No of event	Cum. Survival probability
1	7464	3	0.9996	7525	0	1.000	2573	0	1.0000
2	7437	12	0.9980	7513	18	0.998	2561	4	0.9984
3	7415	2	0.9977	7478	9	0.996	2548	6	0.9961
4	7406	29	0.9938	7453	15	0.994	2539	4	0.9945
5	7374	104	0.9798	7422	2	0.994	2531	1	0.9941

6	7260	4414	0.3841	7417	5	0.993	2523	92	0.9579
7	2813	2282	0.0725	7410	593	0.914	2427	13	0.9527
8	419	330	0.0154	6783	6427	0.048	2413	2361	0.0205

The analysis has been done and it was found that there were 419 remaining diploma students at risk and the proportion of survival was 0.0154. There are 2413 students at risk and 0.0205 proportions of students survive for STPM, and 6783 students at risk and 0.048 proportions of survival remaining at the end of the study for matriculation. The proportion survival of STPM and matriculation qualification recorded the highest probability of surviving in all semester students enrolled except for students enrolled 6 semesters and 8 semesters. Meanwhile, it was found that the proportion of diploma students survives was the lowest for each semester. The log rank test had a chi-squared distribution, χ^2 with two degrees of distribution. The ${\chi^2}_{value}$ = 13,542 obtained from the analysis was higher than ${\chi^2}_{\alpha=0.05}$ = 5.991. It indicated that the qualification entry groups statistically significantly difference in survival. These qualification showed difference survival time when pursues the degree in UTHM at the significance level 0.05. It concluded that there were significant difference between STPM, Diploma and Matriculation curve.

B.Parametric survival analysis: Survival estimates obtained from parametric survival models typically yield plots more consistent with a theoretical survival curve. The advantages of using parametric survival model were the simplicity and completeness. Table 2 shows the results of parametric survival model based on exponential, Weibull and gamma models. The AIC and BIC values from Weibull models in Table 2 recorded the minimum values, therefore Weibull model has been chosen as the more appropriate model to be used in modelling students' performance.

Table 2: Result of fitting parametric survival model to the students achievement

Parametric survival model	Parameter	AIC	BIC	Model
Exponential Model	λ= 0.13201	101190.60	101198.3735	$S(t) = \; e^{-0.13201 \; t}$
Weibull Model	γ = 0.1305 λ =0.0945	43954.70	43970.2469	$S(t) = e^{-0.0945 \; (t^{0.1305})}$
Gamma Model	γ = 49.817 λ = 0.1462	49331.94	49347.4869	$S(t) = 1 - I_{49.8170} \; (0.1462 \; x)$

Further analysis using Weibull model for each of qualifications entry to identify the survival performance of the students. Table 3 presents the λ and γ values of Weibull model. It was found that student with STPM and matriculation more robust or higher survival probability compare to students with diploma qualification.

Table3. The λ and γ values of Weibull model for each qualification

Qualification entry	λ	γ
Diploma	6.7244	0.1060
Matriculation	7.9941	0.0177
STPM	7.9962	0.0172

IV. CONCLUSION

Kaplan Meier techniques well estimated the survival curve of different qualifications based on the semester enrolled by the students. The use of nonparametric survival models greatly reduce cost, a number of sample size and time to follow up which can make the objectives of the study more ethical and achievable. However, this approach cannot be used to estimate or forecast the students' achievement at any time given, *t*. Furthermore, analysis showed that students' performance are significantly different among the qualification entrance to degree programs, namely diploma, matriculation and STPM respectively. While, by using parametric survival model, factor variable can be considered, such as status and time variables in analysing of lifetime distributions. The factor variables in this study were students qualification entry, time variables in the semester enrolled by the students and status variables were students had grade point average greater than 2.00 considered success, meanwhile students' with grade point average less than 2.00 was considered failed during his/her study. The advantages of parametric models include simplicity, availability of likelihood based inference procedures and ease of use of description, comparison, prediction and decision. The most interesting is parametric model can be used to estimate or forecast students' performance at any given time, *t*.The distribution of the data also give a significance contribution in modelling process and modelling performance. Further study should consider or explore other types of distribution that will be more reliable and suitable model.

ACKNOWLEDGEMENT

The authors would like to thank to University Tun Hussein Onn Malaysia, Academic Students' Department for providing us with the secondary data of the student achievement for every semester and valuable source of information. We thank to the editor, associate editor, and the reviewers for their constructive comments that improved the quality of this article. This research was sponsored, in part, by Postgraduate Research Grant (GPPS) U604.

REFERENCES

[1] Zuaini, I., Aziah, A. M. and Rosliza, M. Z. "Factor that determine matriculation students in account subjects". Malaysian Journal of Learning and Instruction (MJLI). 5: pg no: 99-115, 2008.

[2] Norhani, B., Zainab, A. R., Hamidah, A. R. and Aminah, A. K. "Punca prestasi pembelajaran yang lemah di kalangan pelajar Fakulti Pengurusan dan Pembangunan Sumber Manusia UTM". Jurnal Teknologi UTM, 43(E): pg no: 29-44, 2005.

[3] Hafizah, H., Norbahiah, M., Norhana, A., Mimi, D. Z. and Nurhanum, S. S. "Analisis Kuantitatif pencapaian akademik pelajar JKEES". Pascasidang Kongres Pengajaran dan Pembelajaran UKM. pg no: 95-103, 2011.

[4] Mat, U., N., Buniyamin, P. M., Arsad, R. and Kassim. "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention, in: Engineering Education (ICEED)", 2013 IEEE 5th Conference on, IEEE. pg no:126–130, 2013.

[5] Amirah, M. S., Wahidah, H. and Nuraini, A. R. "A review on predicting students' performance using Data Mining techniques", Elsevier.72,pg no: 414-422, 2015.

[6] Ronco, S. L. "Meandering Ways: Studying student dropout with survival analysis", Proceedings of the Association for Institutional Research 34th Annual Forum, New Orleans, LA, 1994.

[7] Tamada, Mike and Inman, C. "Survival Analysis of Faculty Retention Data: How Long Do They Stay?" AIR national conference Orlando FL.1997.

[8] Romero, C. and Ventura, S. "Educational data mining: A review of the state of the art", Trans. Sys. Man Cyber Part C. 40(6), pg no: 601–618, 2010.

[9] Angeline, D. M. D. "Association rule generation for student performance analysis using algorithm", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA).1(1), pg no: 12–16, 2006.

[10] Youngkyoung, M., Guili, Z., Russeli, A. L., Timothy, J. A. and Matthew, W. O. "Nonparametric survival analysis of the loss rate of undergraduate engineering students". Journal of Engineering Educations.100(22), pg no: 349–373, 2011.

[11] Willet, J. B., and Singer, J. D. "From whether to when: new method for studying student dropout and teacher attrition". Review of educational research, 61(4), pg no: 407–450, 1991.

[12] Cox D.R. & Oakes D. "Analysis of survival data". London, Chapman and Hall edition, 1984.

[13] Kalbfleisch, J. D. andPrentice, R. L. "The statistical analysis of failure time data". Wiley Series in Probability and Statistics, New York, 2nd Ed. 2002.

[14] Akram, M., Aman, M. & Taj, R. "Survival analysis of cancer patients using parametric and non-parametric approaches". Journal of Pakistan Vet, 27(4), pg no: 194-198, 2007.

AUTHOR'S BIOGRAPHY

Azme Khamis	Professor, Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia
Che Azmeeza Che Hamat	Research Assistant, Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia
Mohd Asrul Effendi Abdullah	Senior lecturer, Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia
Sabariah Saharan	Lecturer, Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia
Norhaidah Mohd Asrah	Lecturer, Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia