Modeling Students' Performance using Cox and Parametric Accelerated Failure Time Models

Azme Khamis, Che Azmeeza Che Hamat & Mohd Asrul Affendi Abdullah

Department of Mathematics and Statistics
Faculty of Applied Sciences and Technology
University Tun Hussein Onn Malaysia,

DOI: 10.31364/SCIRJ/v8.i7.2020.P0720785
http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785

Abstract: *This study explored the use of survival analysis to investigate the Bachelor's degree students' performance based on GPA, entrance qualification, faculty, and course. The study considered the application of semi-parametric and parametric Accelerated Failure Time (AFT) models. The main objectives of the study are to identify the covariates that dominate students' performance via the Cox model, to investigate the performance of the Cox model based on the Proportional Hazard (PH) assumption, and to compare the performance of parametric AFT models using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Corrected Akaike Information Criterion (AICc). Results revealed that the Cox model suggested the covariates of GPA, faculty, entrance qualification, and the course had affected student performance. PH assumption in the Cox model was violated. This analysis concluded that the Cox model provided a less accurate estimate of student performance and further study should be conducted with parametric AFT models. In parametric AFT models, the Log-normal AFT was chosen as the best model and can be used as an alternative model for estimating student performance at universities and other similar higher educational institutions.*

***Keywords*:** Cox model, Accelerated Failure Time (AFT) model, Log-normal AFT model

1.0 INTRODUCTION

The government subsidizes the entrance of the students who have a chance to receive higher education in a public university in Malaysia. The three entrance requirements to pursue a bachelor's degree at the Institutions of Higher Education (IHE) are Diploma, Matriculation, and the Malaysian Higher School Certificate (STPM). The selection criteria for students who want to pursue a bachelor's degree are under the control of the Ministry of Higher Education (MoHE). The performance of university students is frequently focused on by measuring the learning assessment and curriculum (Mat et al., 2014) or by the final grade earned that is adequate Grade Percentage Average (GPA) achieved (Shahiri, Husain, & Rashid, 2015).

The transition from secondary to tertiary education (university education level) is an extremely stressful experience for most first-year students (Tinto, 1993). As a start, first-year students need to adapt academically and socially to stay in university. Many studies have shown that the difficulty faced by students in adjusting to a new environment in the university can cause them to fail before graduation (Gerdes & Mallinckrodt, 1994). As reported by Martin, Swartz, and Madson (1999), students who have not performed during university studies have a problem adapting to the university's learning style.

Previous research has identified that the factors contributing to students' performance were gender (Zuilkowski & Jukes, 2014), parents' education level (Alarcon & Edwards, 2013), geographical background (Hovdhaugen, 2011), entrance grades of students (Leonavicius, 2009), and GPA (Bruinsma & Jansen, 2009). Among these, GPA was found to be the most critical factor. Students with a lower GPA during the early semesters of the study were less likely to perform well compared to students with higher GPA (Bowers, 2010).

This research determined the efficiency of the Cox model that is commonly used to estimate the performance of students using the survival analysis method. The Cox model identified that the covariates of GPA, entrance qualification, course, and faculty affected student performance. The parametric Accelerated Failure Time (AFT) model serves as a basis for future research in the academic sector. The study can be used to monitor the performance of students in higher education institutions.

2.0 LITERATURE REVIEW

As stated by Morita (2012), survival analysis is a set of statistical techniques that can be used to analyze data where the outcome variable is the time until the occurrence of an event of interest. Survival analysis is commonly used in medical research (Narendranathan & Stewart, 1993) and demographical studies to examine the time until some specified event occurs (Balakrishnan, 1991). The term survival comes from medical research and widely applied in the pharmaceutical sector (Nilsson, 2011). In non-medical contexts, survival analysis is a failure-time analysis, reliability analysis, and event history analysis.

http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785

Cox (1972) proficiently developed the Cox model survival analysis, which derives robust, consistent, and efficient estimates of the covariate effect. As contended by Cox (1972), the Cox model emerged as the most widely used method in survival analysis. The Cox model has the assumption of proportional hazard (PH) rate being constant over time. The results of the study lead to severe bias, wrong inference, or lower power of a test if the PH assumption in the Cox model is not met (Abrahamamowicz, Mackenzie, & Esdaile, 1996).

Patel, Kay, and Rowell (2006) compared the applications of PH and AFT models. The PH model routinely employed for the analysis of time-to-event data. If the PH assumption of violated, the result of the PH model would be difficult to generalize to situations where the length of follow up is different from the one used in the analysis. The study by Swindell (2009) found that the AFT model is an alternative strategy for the analysis of time-to-event data and suitable for use in a further investigation. The AFT model provides a closer examination of the data. As mentioned by Nardi and Schemper (2003), the AFT model is more accessible to interpret and more relevant to the analysis. The AFT model directly analyzed whether the covariates that existed in the study have prolonged or shortened the survival of students during their studies.

3.0 METHODOLOGY

The study was conducted on 2606 students and used the covariates of GPA, entrance qualification, faculty, and course for each student. The study used the survival approach to find the time-dependent variables related to student performance. The event of interest in this study was student failing during their studies with a GPA of less than two. The next term was censored and defined as censored data when at the end of the study, the student did not face an event and had a GPA of more than or equal to two. Students who survived during the study were labeled as performers while those who did not obtain a GPA ≥ 2.00 during the examination considered as non-performers.

3.1 Cox model

The research was carried out using the Cox model. The Cox model is a regression method for survival data. It provides an estimate of the hazard rate that is always non-negative. Hosmer and Lemeshow (1999) presented the mathematical equation of the Cox model which can be written as

$$h(t) = exp\{h_0(t) + a_1x_1 + a_2x_2 + \cdots + a_px_p\} \quad (1)$$

or

$$\log h(t) = h_0(t) + a_1x_1 + a_2x_2 + \cdots + a_px_p \quad (2)$$

where, (Ht) represents the hazard function within the limited period of t and $x_1, x_2, \ldots, _p XP$ are covariates values in the study by groups.

3.2 Proportional Hazard (PH) assumption

The research was carried out to monitor the capability of fitting a statistical model by testing the PH assumption of the Cox model using the Schoenfeld's residuals method. The PH assumption in the Cox model is the hazard rate (HR) is constant over time. The Schoenfeld's residuals method was proposed by Schoenfeld (1982) as a partial residual that is essential to the interpretation of violation of the PH assumption.The Schoenfeld's residuals test was conducted to identify if the PH assumption could be rejected. The PH assumption is not dismissed for a more significant *p*-value (> 0.10) while a smaller *p*-value (<0.10) leads to the rejection of the PH assumption. If the PH assumption violated, then the study should be furthered with the parametric model (Therneau & Grambsch, 2000).

3.3 Accelerated Failure Time (AFT) model

Then, the research was carried out using the parametric model. The AFT model is a technique in the parametric model that is used to incorporate covariates into the survival model. The AFT model differs from the parametric survival model (Exponential, Weibull model, Log-logistic model, and Log-normal model) since the parametric AFT models (Exponential AFT, Weibull AFT, Log-logistic AFT, and Log-normal AFT) can analyze multiple covariates in the analysis (Minh, 2014).

The AFT model was generalized to the situation where failure times of the covariate have recorded for each in the study. The hazard function of the i^{th} individual at time t, $h_i(t)$, is given by Collet (2003) as

$$h_i(t) = e^{-\eta_i}h_0\left(\frac{t}{e^{\eta_i}}\right) \quad (3)$$

Where $\eta_i = \alpha_1x_{1i} + \alpha_2x_{2i} + \cdots + \alpha_px_{pi}$ is the linear component of the model, $x_{1i} + x_{2i} + \cdots + x_{pi}$ is the covariate value in the study by groups, p is the failure time, i is the groups of covariates in the study and t is the time of study

The baseline hazard function, $h_0(t)$, is the hazard of failure at time t for an individual for whom the values of the covariates, x_{ij} are all equal to zero. The corresponding survival function for the i^{th} individual, $S_i(t)$ is

$$S_i(t) = S_0\left\{\frac{t}{exp(\eta_i)}\right\}, \quad (4)$$

where $S_0(t)$ is the baseline survival function.

3.4 Log-linear form of the AFT model

Collet (2003) stated that the AFT model presupposes that the linear function of the covariates serves as the time logarithm. The formula has been presented by Collet (2003) as:

http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\epsilon_i \quad (5)$$

Consider a Log-linear model for the random variable, T_i associated with the lifetime of the i^{th} individual in a survival study. In this model, $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the unknown coefficients of the explanatory variables $x_{1i}, x_{2i}, \ldots x_{pi}$, while μ and σ are the intercept and scale parameter, respectively. ϵ_i refers to the random variable used to model the deviation of the values of log T_i from the linear part of the model, and ϵ_i is assumed to have a specific probability distribution. The survival function for the i^{th} individual, $S_i(t)$ is

$$S_i(t) = \mathrm{P}\left(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\epsilon_i \geq \log t\right),$$
$$= P\left(\epsilon_i \geq \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \quad (6)$$

$S_{\epsilon_i}(\epsilon)$ is the survival function of the random variable ϵ_i in the Log-linear model of Equation (5). The survival function of the i^{th} individual in Equation (6) can be expressed as

$$S_i(t) = S_{\epsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \quad (7)$$

The parametric AFT model was employed in this study to identify the best alternative model to propose in the survey of student performance. The performance of the parametric AFT model was compared using AIC, BIC, and AIC_C values to identify the best model. Then, the time ratio (TR) of student performance following the best model in this study.

4.0 EMPERICAL RESULT

The Cox model was applied to identify the covariates that dominate student performance. The study also investigated the performance of the Cox model by testing the PH assumption using Schoenfeld's residuals.

The available covariates included in the study. The selection of covariates broadly depended on the AIC and *p*-value. The *p*-value was measured to keep or remove the covariates from the model. The covariates can be retained in the model when the *p*-value is significant at 95% significance level, or *p*-value is less than 0.05. The results show in Table 1.

Table 1 Covariates selection in the Cox PH model

Covariates in model	AIC		*p*-value
GPA + Qualification + Course + Faculty	11304.14	GPA	0.0000
		Qualification	0.0000
		Course	0.0000
		Faculty	0.0398

All covariates are significant at the 95% significance level. AIC values are small when the model has covariates of GPA, qualification, course, and faculty. Hence, all covariates are significant to included in the Cox model analysis. After the covariates selection, the Cox model checking conducted by using Schoenfeld's residuals at the significance level of 90%. Figure 1 below shows that the scaled Schoenfeld's residuals correlated with each other and the separate plots are not around zero. It means there is a correlation between each covariate of the Cox model and time, which implies that the PH assumption violated. To minimize the subjectivity of the graphical method, a formal test for checking the covariates tabulated in Table 2.

http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785

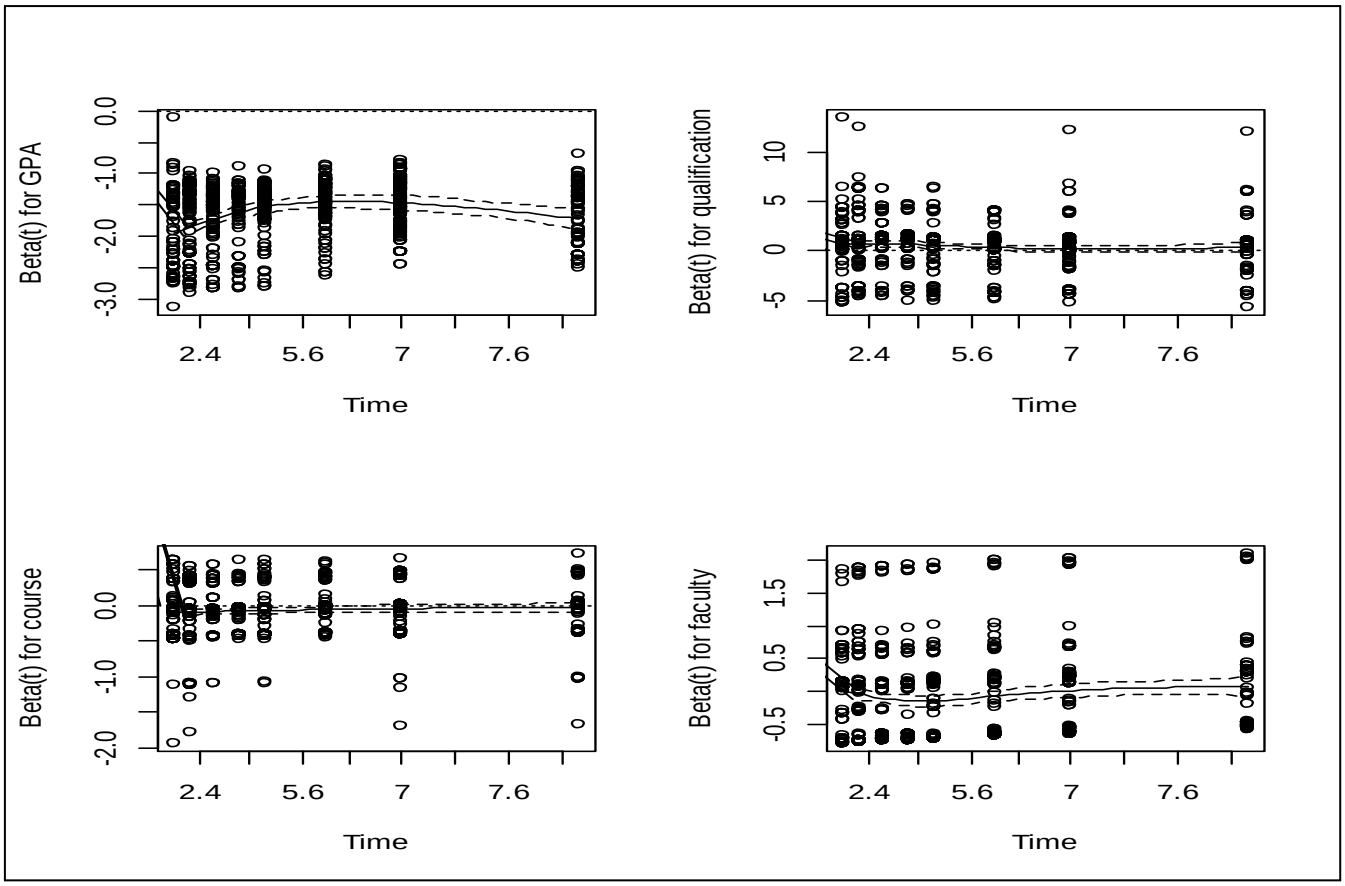

Fig 2 Scaled Schoenfeld residuals plots for each covariate versus semester in model fit to the student data.

Table 2 shows that χ^2 values are large and all the *p*-values are less than 0.10 at the 90% significance level. Hence, the null hypothesis rejected. It means there is a correlation between each covariate of the Cox model. Schoenfeld's residuals analysis shows that the performance of the Cox model is less accurate due to the PH assumption violation. A further study should be conducted using other survival regression models such as the parametric AFT models.

Table 2 Scaled Schoenfeld's Residuals of Covariates on the PH.

Covariate	χ^2	*p*-value
GPA	5.20	0.0225
Qualification	3.23	0.0723
Course	3.11	0.0776
Faculty	3.72	0.0539
All covariates	10.94	0.0273

The study proceeds with the parametric AFT model at 99% significance level. The *p*-values for all covariates in each model are 0.0000. This result shows that the Exponential AFT, Weibull AFT, Log-logistic AFT, and Log-normal AFT models fit the data well with all covariates giving a significant effect to the study.

Table 3 shows the AIC, BIC, and AIC_C values for each parametric AFT model. The Log-normal AFT model is the best parametric AFT model based on the highest Log-likelihood value and smallest values of AIC, BIC, and AIC_C criteria.

Table 3 Log-likelihood, AIC, AIC_C, and BIC values for parametric AFT models.

Model	Log-likelihood	AIC	BIC	AICc
Exponential AFT model	-2963.3000	5936.6950	5947.8226	5936.6034
Weibull AFT model	-2657.2000	5326.4210	5339.8671	5326.4048
Log-normal AFT model	-2461.0000	4934.0970	4947.4671	4934.0048
Log-logistic AFT model	-2463.4000	4938.7450	4952.2671	4938.8048

Table 4 shows the estimates, TR, and *p*-values for the Log-normal AFT model. The estimated values of covariates using the Log-normal AFT model are 0.9059 for GPA, -0.1671 for qualification, 0.0265 for course, and 0.0313 for faculty. These results show the estimated values of covariates have significant use in determining the survival of Bachelor's degree students in UTHM,

http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785

as the *p*-value is less than 0.01. For simplicity and ease of interpretation, TR reported. TR is estimated to find out the trend of covariates affecting the survival time among UTHM students.

In Table 4, the TR for GPA is 2.4741, courses are 1.0268, and faculty is 1.0318. All TR values are more significant than one. It concluded that GPA, classes, and faculty covariates effected on the increase in survival time among students. The TR for entrance qualification covariate is 0.8461, which is less than one. It concluded that this covariate an effected on the decrease in the survival time among undergraduate students in UTHM or this covariate indicates an earlier time to the event occurrence or failure rate of 84%.

Table 4 The Estimate, Time Ratio (TR), and *p*-value for the Log-normal AFT model.

Covariate	Estimate	Time ratio	*p*-value
GPA	0.9059	2.4741	0.0000
Qualification	-0.1671	0.8461	0.0000
Course	0.0265	1.0268	0.0000
Faculty	0.0313	1.0318	0.0000

5.0 DISCUSSION AND CONCLUSION

The most popular method for examining the effect of multiple covariates in the Cox model. The covariates used in this study are GPA, qualification, course, and faculty. All of the covariates used in this study are significant at the 95% significance level and are meant to include in this study. After the covariates selection, the Cox model performance checking was conducted using Schoenfeld's residuals at the significance level of 90%. The analysis found the Cox model was not able to give an accurate result for the study since the PH assumption analyzed by Schoenfeld's residuals in the Cox model was violated with *p*-values less than 0.10. As the Cox model result could be invalid, the parametric model analysis was conducted to produce more accurate results.

The performance of the parametric AFT models was compared to identify the best alternative model to propose in the study on student performance as an alternative to the Cox model. Various parametric AFT models applied in this analysis, namely the Exponential AFT model, Weibull AFT model, Log-normal AFT model, and Log-logistic AFT model. Results revealed that the Log-normal AFT model gave the best fit based on the lowest AIC, BIC, and AIC_C values. This model comparison suggested that the Log-normal AFT model is the best alternative model to study students' performance further.

Although the Cox model with PH assumption has come to the fore in most research in the education field, results of the Log-normal AFT model have often been more valid and have a minor bias since this model has a better fit due to specific statistical distribution for the survival time (Zare et al., 2015). The Log-normal AFT model is also a reliable alternative to the Cox model(Nardi & Schemper, 2003). The Log-normal AFT model offers some benefits that lead to the more efficient estimation of risk of failure and survival of students than the Cox model used in this study. Since the PH does not follow an assumption is in the Log-normal AFT model, the model can give some insights to the research on student performance.

Acknowledgment

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

6.0 REFERENCE

Abrahamowicz, M., Mackenzie. T. & Esdaile, J. M. (1996). Time-dependent Hazard Ratio: Modeling and Hypothesis Testing with Application in Lupus Nephritis. *Journal of American Statistical Association*, Vol. 91. No. 433 pp. 1432–1439.

Alarcon, G.M. & Edwards, J.M. (2013). Ability and Motivation: Assessing Individual Factors That Contribute to University Retention. *Journal of Educational Psychology*, Vol. 105. No. 1 pp. 129-137. https://sci-hub.tw/10.1037/a0028496

Balakrishnan, N. (1991). *Handbook of the Logistic Distribution*. Hamilton: CRC Press.

Bowers, A. G. (2010). Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts. *The Journal of Educational Research*, Vol. 103. No. 3 pp. 191-207.

Bruinsma, M. & Jansen, E.P.W.A. (2009). When will I Succeed in My First-Year Diploma? Survival Analysis in Dutch Higher Education. *Higher Education Research & Development*, Vol. *28*. No. 1 pp. 99-114.

Collett, D. (2003). *Modeling Survival Data in Medical Research*. 2nd ed. London: Chapman and Hall.

Cox, D.R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, Vol. 34. No. 2 pp.187-220.

Gerdes, H. & Mallinckrodt, B. (1994). Emotional, Social, and Academic Adjustment of College Students: A Longitudinal Study of Retention. Journal of Counseling and Development, Vol. 72. No. 3 pp. 281-288. https://sci-hub.tw/10.1002/j.1556-6676.1994.tb00935.x

Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis*. 1st ed. New York: John Wiley and Sons.

Hovdhaugen, E. (2011). Do Structured Study Programmers Lead to Lower Rates of Dropout and Student Transfer from University? *Irish Educational Studies*, Vol. 30. No. 22 pp. 237-251. https://sci-hub.tw/10.1080/03323315.2011.569143

Leonavicius, G. (2009). Research into Bachelor's Degree Studies of Informatics at VPU using Survival Analysis Methods. *Pedagogical Studies*, Vol. 94 pp. 95-98.

http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785

Martin, W.E., Swartz-Kulstad, J.L. & Madson, M. (1999). Psychosocial Factors That Predict the College Adjustment of First-Year Undergraduate Students: Implications for College Counsellors. *Journal of College Counselling*, Vol. 2. No. 2 pp. 121-133. https://sci-hub.tw/10.1002/j.2161-1882.1999.tb00150.x

Mat, U., Buniyamin, N., Arsad, P.M., & Kassim, R.A. (2014). An Overview of using Academic Analytics to Predict and Improve Students' Achievement: A Proposed Proactive Intelligent Intervention. In *2013 IEEE 5th International Conference on Engineering Education: Aligning Engineering Education with Industrial Needs for Nation Development, ICEED 2013*, pp. 126-130.

Minh, H. P. (2014). Survival Analysis: Breast Cancer. *Undergraduate Journal of Mathematical Modelling: One + Two*, Vol. 6. No. 4 pp. 1-20. https://pdfs.semanticscholar.org/fc62/b5366f849c7d7c895f5cc50de08e036aaf6c.pdf

Morita, J.G., Lee, T.W. & Mowday, R.T. (2012). The Regression-Analog to Survival Analysis : A Selected Application. *The Academy of Management Journal*, Vol. 36. No. 6 pp. 1430–1464.

Nardi, A. & Schemper, M. (2003). Comparing Cox and Parametric Models in Clinical Studies. *Statistics in Medicine*, Vol. 22. No. 23 pp. 3597-3610. https://pdfs.semanticscholar.org/88af/8f99b4870dd374f73cf17503217ca0bd072d.pdf

Narendranathan, W & Stewart, M. (1993). Modeling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Baseline Hazards, *Journal of the Royal Statistical Society*, Vol. 42. No. 1 pp. 63-83.

Nilsson, M. (2011). A Survival Analysis of Fixation Times in Reading. *Proceeding CMCL '11 Proceedings of the Second Workshop on Cognitive Modelling and Computational Linguistic.* Uppsala University. pp. 107-115.

Patel, K., Kay, R. & Rowell, L. (2006). Comparing Proportional Hazards and Accelerated Failure Time Models: An Application in Influenza. *Pharmaceutical Statistics*, Vol. 5. No. 3 pp. 213–224. https://sci-hub.tw/10.1002/pst.213

Shahiri, A.M., Husain, W. & Rashid, N.A. (2015). A Review on Predicting Student's Performance using Data Mining Techniques. *Elsevier*, Vol. 72 pp. 414 – 422. https://www.sciencedirect.com/science/article/pii/S1877050915036182

Schoenfeld, D. (1982). Partial Residuals for the Proportional Hazards Regression Model. *Biometrika*, Vol. 69 pp. 239 – 241.

Swindell, W. R. (2009). Accelerated Failure Time Models Provide a Useful Statistical Framework for Aging Research. *Experimental Gerontology Journal*, Vol 44. No. 3 pp. 190-200.

Therneau, T. M. & Grambsch, P. M. (2000). *Modelling Survival Data: Extending the Cox Model*. New York: Springer.

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition.* 2nd ed. Chicago: University of Chicago Press.

Zare, A., Hosseini, M., Mahmoodi, M., Mohammad, K., Zeraati, H., & Holakouie Naieni, K. (2015). A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients. *Iranian Journal of Public Health*, Vol. 44. No. 8 pp. 1095–1102.

Zuilkowski, S.S. & Jukes, M.C.H. (2014). Early Childhood Malaria Prevention and Children's Pattern of School Leaving in The Gambia. *British Journal of Educational Psychology*, Vol. 84 pp. 483-501.

http://dx.doi.org/10.31364/SCIRJ/v8.i7.2020.P0720785