Modelling lecturer performance index of private university in Tulungagung by using survival analysis with multivariate adaptive regression spline

M Hasyim\(^1\) and D D Prastyo\(^2\)

\(^1\)Study Program of Mathematics Education, STKIP Tulungagung, Indonesia
\(^2\)Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail: \(^1\)maylita@stkippgritulungagung.ac.id, \(^2\)dedy-dp@statistika.its.ac.id

**Abstract.** Survival analysis performs relationship between independent variables and survival time as dependent variable. In fact, not all survival data can be recorded completely by any reasons. In such situation, the data is called censored data. Moreover, several model for survival analysis requires assumptions. One of the approaches in survival analysis is nonparametric that gives more relax assumption. In this research, the nonparametric approach that is employed is Multivariate Regression Adaptive Spline (MARS). This study is aimed to measure the performance of private university’s lecturer. The survival time in this study is duration needed by lecturer to obtain their professional certificate. The results show that research activities is a significant factor along with developing courses material, good publication in international or national journal, and activities in research collaboration.

1. **Introduction**

Survival analysis is one of the statistical methods that are specially used to process the data or cases related to the time duration until an event occurs. The typical data in survival analysis is the existence of censored data [1]. At the beginning of its development, the study about survival is concerned on the probability prediction of average life expectancy. Recently, survival analysis has been developed for identification of risk factors and prognostic factors associated with the maturity of the disease [2] and many other areas.

The survival analysis describes a process that is associated with time, from the time origin until the occurrence of a special occurrence or end point [3]. There several approaches in survival analysis, for example parametric and nonparametric approaches. One of the nonparametric approaches that can be employed is Multivariate Adaptive Regression Splines (MARS) as a nonparametric and semi parametric regression that takes into accounts the covariates with a multivariate approach [4].

Nonparametric regression using MARS does not depend on the assumption of a certain curve shape, has flexibility in high dimensional data, and modelling involves a lot of interaction with a few variables [5]. In survival analysis using MARS approach, the residuals of Cox Proportional Hazard Model (Cox PHM) is treated as dependent variable in MARS modelling [6]. Reference [7] is an example of research that employ survival analysis using MARS approach in Dengue fever cases, where the dependent variable of MARS models is martingale residual for uncensored data. Another study that employs Cox PHM with MARS approach also had been done for the case of...
survival of heart attack patients in Germany. It shows that the MARS approach gives better results than Cox PHM [6]. Reference [8] employed such approach in digital economic case.

Another nonparametric approach for survival analysis is survival support vector machine, i.e. the modification of support vector machines (SVM) for survival analysis modelling. References [9, 10] developed additive survival SVM and applied it in medical cases. A simulation studies on survival SVM had been done by [11, 12] and applied to cervical cancer study. The survival SVM and survival MARS have different approaches.

In this study, the lecturer performance index or Indeks Kinerja Dosen (IKD) is modelled by using survival MARS. Lecturers are professional educators and scientists with the main task of transforming, developing and disseminating science, technology and the arts through education, research and dedication to society. Lecturer performance is defined as the ability to carry out the work or tasks that lecturers have in completing a job. Performance can be interpreted as work presentation, job performance, work achievement, or work result. It also refers to the result or output of a process. Education is aimed to: (1) improve the performance, capability and output of education; (2) facilitate communication and exchange of information about best practice of education with various types of educational institutions; and (3) as tools to improve institutional performance education and guidelines in strategic planning. By using Survival MARS, it will be known significant factors that influence the time needed by lectures to get their professional certificate.

2. Literature review
2.1. Survival analysis
Survival analysis is a statistical method that can be used to analyse data that related to start time (time origin) or start point until the specific event happened (end point) or failure event [13]. To determining the survival time, there are three term required: (i) Time origin (starting point), the starting time to record, (ii) Ending event of interest (recent events), the time when a lecture got the professional certificate, (iii) Measurement scale for the passage of time as a limit of the time of incident from the beginning to the end. The scale is measured in days, weeks, months, or years. In this study measuring scale used the time duration when a lecture was in years.

In survival analysis, there is difficulty in observing data is the possibility of some individual which cannot be observed from the start point to the end point. This situation is called the censored data [1] caused by three reasons: (i) Loss to follow up, it occurs when a lecture decides to move another university, (ii) Drop Out, it occurs at a lecture chooses to pension, (iii) Termination of study, it occurs when the research period was ended while the lecture has not get professional certificate yet.

2.2. Hazard function and survival function
There are two main functions in survival analysis. They are the survival function and hazard function. Survival function is the basis of survival analysis that refers to an individual opportunity to survive over time $t$ [2]. Suppose $T$ is survival time and has density function $f(t)$, the cumulative distribution is denoted as:

$$F(t) = P(T < t) = \int_{0}^{t} f(u) du$$

The survival function $S(t)$ is probability of individual life longer than time $t$ as follows:

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t)$$

The hazard function for estimating individual probability get an event at time $t$ is described as [1].

$$h(t)dt = Pr(t \leq T < t + \Delta t \mid T \geq t)$$

Relationship between density function, hazard function, and survival function is:

$$h(t) = \frac{f(t)}{S(t)}$$

The relationship of cumulative hazard and survival function is:
\[ S(t) = \exp(- \int h(t) \, dt) = \exp(-H(t)) \text{ and } H(t) = -\ln S(t) \] (5)

### 2.3. Cox proportion hazard model

The Cox PHM is proposed by Cox in 1972 [14] and now is widely developed. This modelling is a log-linear relationship between independent variables and hazard function as follows:

\[ h(t \mid X = x) = h_0(t) \exp(\beta'x) \] (6)

where \( h_0(t) \) baseline hazard which is the form is not specified. The baseline hazard has specific form corresponding to the distribution of survival time.

Distribution approximation used in this study is time needed (started from first time to be a lecturer) by lecturer to get professional certificate. The distribution of survival time can be tested by using minimum Anderson Darling score as follows.

\[ A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)[\ln(F(t_i)) + \ln(1 - F(t_{n+1-i}))] \] (7)

### 2.4. Weibull distribution with three parameters

Given the survival time follows Weibull distribution with three parameters; the baseline hazard has specific form accordingly. The cumulative distribution for Weibull distribution with three parameters is:

\[ F(t) = 1 - \exp\left\{-\left(\frac{t-\gamma}{\eta}\right)^\beta\right\}, \ t \geq \gamma \] (8)

with \( \gamma \) is location parameter, \( \eta \) is scale parameter, \( \beta \) is shape parameter, where \( \eta, \beta > 0 \) and \( \gamma \leq t < \infty \). The baseline hazard for survival time that Weibull distribution with three parameters is:

\[ h_0(t | \gamma, \eta, \beta) = \frac{\beta(t-\gamma)^{\beta-1}}{\eta^\beta} \] (9)

### 2.5. Multivariate adaptive regression splines (MARS)

Multivariate Adaptive Regression Spline (MARS) is one of the flexible methods for modelling high-dimensional regression data. MARS is a form of extension of the Basis Spline Functions where the number of basis function is the parameters of the model. Some terms that need to be considered in MARS is as follows,

a. **Knots**, the point of a regression line to form a region of a regression function.

b. **Basis Function (BF)**, collection of some of the functions that are used to describe the relationship between the response variable and the predictor.

c. **Interaction**, a correlation between variables and the maximum number of interaction (MI) 1, 2, and 3.

MARS model is formulated as follows:

\[ f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_k \prod_{k=1}^{K_m} S_{km}(x_{(i,m)} - \tau_{km}) + \varepsilon \] (10)

The best model produces smallest GCV which is denoted as:

\[ GCV(m) = \frac{\text{ASR}}{n - C(M)} = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - f_M(x_i) \right]^2 \left[ 1 - \frac{C(M)}{n} \right] \] (11)

where \( C(M) = \text{trace}(B(B^TB)^{-1}B^T) + 1 \).
3. **Research methodology**
Dependent variable of this study is lecturer performance index (IKD). The IKD is measured for both lecturer as civil servant (DPK) and lecturer with permanent status (DTY). Categorical dependent variable has goal to be suitable with MARS binary response. The data for survival function in this study is the time needed by lecturer as DPK or DTY in private university. The origin point is the starting time becomes lecture. The predictors involve education activities (6 indicators), research activities (4 indicators) and public services (3 indicators). Steps of analyses in this study are as follows:

a. Determining the survival data that will be used and eliminate the censored data.
b. Describing the characteristics of lecturer as research object
c. Fitting the distribution of survival data based on smallest score of Anderson-Darling test
d. Determining the form of baseline hazard function
e. Estimating the survival function and cumulative hazard function
f. Modelling survival function with MARS approach

4. **Analysis and discussion**
The first step of analysis is to fit the distribution of survival data. According to minimum Anderson Darling as reported in Table, the survival time follows Weibull distribution with three parameters such that the baseline hazard is specified accordingly.

| No. | Distribution          | AD Score |
|-----|----------------------|----------|
| 1   | 2-Parameter Weibull  | 1.094    |
| 2   | 2-Parameter Lognormal| 1.760    |
| 3   | Exponential          | 9.235    |
| 4   | Log logistic         | 2.043    |
| 5   | 3-Parameters Weibull | 1.020    |
| 6   | 3-Parameters Lognormal| 2.703  |
| 7   | 3-Parameters Exponential| 11.492 |
| 8   | 3-Parameters Log logistic| 1.185  |

The baseline hazard function is estimated as follows:

\[
h_b(t \mid \hat{\gamma}, \hat{\eta}, \hat{\beta}) = \frac{\hat{\beta}(t - \hat{\gamma})^{\hat{\beta} - 1}}{\hat{\eta}^{\hat{\beta}}} = \frac{3.1503(t - (-1.708))^{6.1503 - 1}}{3.345^{6.1503}}
\]

\[
= 0.00365(t + 1.708)^{5.1503}
\]

According to the baseline hazard function, the calculation of survival function and cumulative hazard function can be made. Estimation of survival function and cumulative hazard are shown on Table 2.

| \(t\) | \(S(t)\)   | \(H(t)\)   |
|-------|-------------|-------------|
| 1     | 0.998105    | 0.001897    |
| 2     | 0.925461    | 0.077463    |
| 3     | 0.462685    | 0.770708    |
| 4     | 0.016751    | 1           |
| 5     | 2.57E-07    | 1           |
Table 2 indicates the probability a lecture get their professional certificate through time $t$. In general, rate of getting certificate is increase in sequence time. A lecturer will get his/her existence of profession between three and four years, starting from first time as a lecture.

Based on the results of trial and error combination BF, MI, and MO, the combination of which produces minimum GCV value is a combination of 22, 3, 1 with a value of GCV = 0.373 with $R^2 = 0.829$. Based on the results of this combination, it is known MARS models produced are as follows:

$$\hat{y} = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[ s_{km}(x_{v(km)}-t_{km}) \right]^2 + \varepsilon_i$$

$$= 0.617 + 0.872BF_6 + 0.539BF_7 + 0.463BF_9 + 0.312BF_{16}$$

and

$$h(t) = h_0(t) \exp(\hat{y})$$

$$h(t) = 0.00365 (1 + 1.708)^{5.1503} \exp(0.617 + 0.872BF_6 + 0.539BF_7 + 0.463BF_9$$

$$+ 0.312BF_{16})$$

where $BF_6$ = research activities, $BF_7$ = submission of article to International/national journal, $BF_9$ = attending conference activities, and $BF_{16}$ = participating in workshop and training.

The results of survival MARS modelling show that in general, the variables that affect the survival to be professional lecturer are research activities, submission International/national journal, Attend conference activities, workshop and training.

5. **Conclusion**

Based on the empirical results using Cox PHM MARS approach, which used a combination Basis Functions, Maximum interaction are 22, 3, and 1 with a minimum GCV value was 0.022. Significant predictors influencing the survival time until the professional certificate for lecturer is obtained are research activities, submission to International/national journal, attending conference activities, and participation in workshop and training. Research activities have the largest contribution in the model, followed by submission to international/national journal, attending the conference, and participating in workshop and training.

**Acknowledgments**

Authors wishing to acknowledge assistance or encouragement from research partner for any advice and idea. Special thanks to Indonesian ministry for research and higher education for the financial support under PEKERTI scheme.

**References**

[1] Kleinbaum, D.G. 2012. *Survival Analysis*. London: Springer

[2] Lee, E.T. 2003 *Statistical Method for survival Data Analysis*. London: John Willey

[3] Collet, D. 2003. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

[4] Friedman, J.H. 1991. Estimating Function of Mixed Ordinal and Categorical Variables Using Multivariate Adaptive Regression Splines. Technical Report LCS 107, Statistics Department, Stanford University

[5] Friedman, J.H. 1991. Multivariate Adaptive Regression Spline. *The Annals of Statistics*, 19: 1-141.

[6] Kriner, M. 2007. *Survival Analysis with Multivariate Adaptive Regression Splines*. Dissertation. Munich University.
[7] Nisa’, F.S. and Budiantara, I.N. 2012. Analisis Survival dengan Pendekatan Multivariate Adaptive Regression Spline pada Kasus Demam Berdarah Dengue (DBD). *Jurnal Sains dan Seni ITS*, 1 (1): 318-323.

[8] Irwansyah, E., Nyoman. D.A. and Bekti R.D. 2015. Cox Proportional Hazard with Multivariate Adaptive Regression Spline to Analyse the Product Sales Time in E-Commerce. *International Journal of Applied Mathematics and Statistics*, 53(5).

[9] Van Belle, V., Pelckmans, K., Suykens, J.A. and Van Huffel, S. 2010. Additive Survival Least-Squares Support Vector Machines, *Statistics in Medicine*, 29(2): 296-308.

[10] Van Belle, V., Pelckmans, K., Suykens, J.A. and Van Huffel, S. 2011. Support Vector Methods for Survival Analysis: A Comparison between Ranking and Regression Approaches, *Artificial Intelligence in Medicine*, 53(2): 107-118.

[11] Khoiri, H.A., Prastyo, D.D. and Purnami, S.W. 2017. A Simulation Study and Application of Feature Selection on Survival Least Square Support Vector Machines. *Proceeding of the 13th IMT-GT International Conference on Mathematics, Statistics, and their Applications (ICMSA2017)*, AIP Conference Proceedings 1905 (050024).

[12] Khotimah, C., Purnami, S.W., Prastyo, D.D., Chosuvivatwong, V. and Sriplung, H. 2017. Additive Survival Least Square Support Vector Machines: A Simulation Study and Its Application to Cervical Cancer Prediction. *Proceeding of the 13th IMT-GT International Conference on Mathematics, Statistics, and their Applications (ICMSA2017)*, AIP Conference Proceedings 1905 (050024).

[13] Le, C.T. 1997. *Applied Survival Analysis*. New York: John Wiley & Sons, Inc.

[14] Cox, D. R. 1972. Regression Model and Live Tables (with discussion), *Journal of The Royal Statistical Society*, 34: 187-220.