M – 07

Cox Proportional Hazard Model with Multivariate Adaptive Regresion Spline

Hendra Dukalang[1], B. W. Otok[2], Ismaini Zain[2], Herlina Yusuf[3],
[1]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[2]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[3]Dept. of Public Health, Universitas Negeri Gorontalo
hendra37.hd@gmail.com

Abstract— Events related to the survival time always happens in everyday life, one of which is time duration that need to recover from illness. Time that we need until the event happened is called survival data. Generally, not all of survival data can be observed and it is called data censored. One of statistical method that can be used to analyze and determine the survival rate of survival data is the cox proportional hazard models. In its development, the residuals of the cox proportional hazard (Cox PH) model can be used as response variable for regression function. The relationship between response variable and predictor variables often is not known the function of regression. So we are needed nonparametric regression. One of method nonparametric regression that can be used is Multivariate Regression Adaptive Spline (MARS). In this study, survival analysis is focused on the patients of HIV/AIDS which is a deadly disease. To determine survival rate of HIV/AIDS patients is used a hazard function and survival function with time duration patient stayed as variable. To know the other factors of the survival of HIV/AIDS patient is used Cox PH Models with MARS approach. The results showed that gender is one factor in the survival of HIV/AIDS patients, and treatment compliance, employment status, CD4 count, age and educational level.

Keywords: *Survival Analysis, Cox PH, MARS, HIV/AIDS*

I. INTRODUCTION

Survival analysis is a statistical analysis that is specifically used to analyze the data or cases related to the time duration until the event happened and there are data censored [1]. At first time, studied of survival is focused on the probability predictions of response, survival, average life expectancy and comparing the treatment of survival illustration experiment in humans. But survival analysis developed in the identification of risk factors and prognostic factors associated with the development of the disease [2]. One method of analysis that can be used for survival data are cox proportional hazards regression (Cox PH). Cox PH regression modeling can also be used to determine which combination of independent variables that influence in the model. In its development, Cox PH regression modeling can include relationships between predictor variables with the model function multivariate regression adaptive spline (MARS).

MARS is one of nonparametric regression method that does not depend on the assumption of a certain curve shape so it has flexibility in high dimensional data and modeling involves a lot of interaction with a few variables [3]. The variable responses in MARS modeling can use the residuals of the modeling Cox PH, so the survival modeling of MARS can be interpreted as MARS modeling the response variable is the residual result of modeling Cox PH [4]. The Previous research has been done to use of survival analysis with MARS approach in DBD cases, where the response variable of MARS models use *martingale residual* for uncensored data [5]. Then Cox proportional hazard and MARS used to analyze product sales with a electronic media system [6]. Previously, they had done research on survival analysis using MARS approach for the case of survival of heart patients in Germany, and show that the MARS method give better results than Cox PH regression [4].

In this study, the Regression Cox PH using MARS approach is used to determine the factors that influence survival of HIV/AIDS patients. Human Immunodeficiency Virus (HIV) is a virus that decrease the body's immune system so that the people affected by this virus will be susceptible to various infections and then causes *Acquired Immune Deficiency Syndrome* (AIDS). Research on HIV/AIDS in Indonesia is more emphasis on efforts to reduce the incidence of HIV/AIDS and how the healing response of

HIV/AIDS. One of them is a mixture survival modelling for HIV/AIDS cases in Semarang [7]. To determine the factors that affect the survival of HIV/AIDS.

II. LITERATURE REVIEW

A. Survival Analysis

Survival analysis is a statistical method that can be used to analyze data that related to start time (time origin) or start point until the specific event happened (end point) or failure event [8]

To determining the survival time, there are three factors required:

1. *Time origin* (starting point), is time to record and analyze an incident when the patients were first declared HIV/AIDS.
2. *Ending event of interest* (recent events) is the expired recording time. This time is useful to know the status of censored or not censored patient to be able to do analysis. Recent events in this study is the time when the HIV/AIDS patients were declared dead.
3. *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 the patients were suffering HIV/AIDS in months.

In survival analysis, there is difficulty data observing that is the possibility of some individual observations who cannot be observed from the start point to the end point, this situation is called the censored data [1]. In this study, there are three causes of censored data.

1. *Loss to follow up*, occurs when the patient decides to move another hospital or refuse to observe.
2. *Drop Out*, occurs the patient chooses to go home.
3. *Termination of Study*, occurs when the research period was ended while the patient has not reached the failure event.

B. Hazard Function and Survival Function

In survival analysis, there are two main functions that is survival function and hazard function [8]. Survival function is the basis of survival analysis, because it includes the probability of survival from the time varying provide important information about the survival data. Survival Function is an individual opportunity who can survive over time *t* [2], and usually denoted by.

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)\,du \tag{1}$$

$$S(t) = \exp\left[-\int_0^t \lambda(u)\,du \right] \tag{2}$$

Hazard Function is an individual probability to reach specific incidents at time interval $(t, \Delta t)$ with individual assuming to stay on at this time interval. And usually denoted by $\lambda(t)$. This function is used to express the *hazard rate* or the rate of cure and survival up to time-*t*.

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{3}$$

Where $f(t)$ is probability density function (PDF) on the distribution of the estimated survival data, and it is known that:

$$\int_0^t f(u)\,du = 1 - S(t) \tag{4}$$

So generally, the relationship of survival function and cumulative hazard function based on that equation is as follows:

$$\Lambda(t) = -\ln\ \mathrm{S}(t) \tag{5}$$

C. Distribution Estimates

Estimation of distribution used to the survival data which in this study is duration of suffering HIV/AIDS patients to otherwise experience *failure event*. Estimation of distribution is conducted by

Anderson-Darling test (AD) because it has a strong strength and accurate if we compared with other distribution test [9].

Equation Anderson-Darling test statistic (AD) is as follows:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2_i - 1)\left[\ln F(Y_i) + \ln F(Y_{n+1-i})\right] \quad (6)$$

Where: F = the cumulative distribution function of the conjecture distribution.
Y = survival time data.
n = number of sample

D. Cox Proportional Hazard Model

Regression modeling to determine the factors that influence survival data for uncensored data is called Cox Proportional Hazard Regression models [10]. Cox PH regression is used when the observed outcome was the length of time of an event. This Modeling is a log-linear relationship between X and the general function of hazard on T are as follows:

$$\lambda(t|X - x) = \lambda_0(t)e^{\beta x} \quad (7)$$

For variable X that has covariate, the equation used is as follows:

$$\lambda_i(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} \quad (8)$$

Where:

$\lambda_i(t)$ = hazard function for individual to $-i$

$\lambda_0(t)$ = baseline hazard

$\beta_1, \beta_2, \ldots, \beta_p$ = coefficient regression

$x_1, x_2, \ldots, x_p$ = variable value for individual to $-i$

The most important assumptions that must be met in the regression is Cox Proportional Hazard assumptions which means that the ratio of the hazard function is constant over time or equivalent to the statement that the ratio of the hazard function of an individual against another individual hazard function is proportional. This research will use the approach chart using log minus log survival plots to check the assumptions Proportional Hazard. According to the Cox regression model, the hazard function for failure individual-i for time-t can be written as in Equation (9) is as follows:

$$\lambda_i(t) = \lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j\right) \quad (9)$$

Modelling using Cox Proportional Hazard produces two types of residual, that is Martingale Residual and Deviance Residual that obtained from Cox Null Model. This study used Martingale Residual which serves as the response variable for modeling MARS. Residual Martingale equation is as follows:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds \quad (10)$$

$$M_i(t) = N_i(t) - \hat{\Lambda}_i(t) \quad (11)$$

Where

$M_i(t)$ = Martingale Residual-i at time-t

$N_i(t)$ = The process of counting events (data uncensored given value of 1 and data censored given value of 0) for data-i at time-t

$Y_i(s)$ = Indicators, if subject-i is *under risk immediately* before-t

$\hat{\Lambda}_i(t)$ = Breslow estimator of the cumulative baseline hazard function

E. Multivariate Adaptife Regression Spline

Multivariate Adaptive Regression Splines (MARS) is one of the new flexible method for modeling high-dimensional regression data. MARS is a form of extension of the Basis Splines Functions where the number of basis function is the parameters of the model.

Some terms that need to be considered in the methods and modeling MARS is as follows,

1. *Knots* is the point of a regression line to form a region of a regression function.
2. *Basis Function* (BF) is a collection of some of the functions that are used to describe the relationship between the response variable and the predictor variable.
3. Interaction is a correlation between variables and the maximum number of interaction (MI) 1, 2, and 3.

The general equation MARS models are as follows:

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - \tau_{km} \right) \right] + \varepsilon \tag{12}$$

Estimator model of multivariate adaptive regression splines or MARS [3]:

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - \tau_{km} \right) \right] + \varepsilon \tag{13}$$

Where the first summation covers all the bases for a single variable functions, covering all the bases the second summation function for the interaction between two variables, the third summation includes all the base functionality for the interaction between the three variables and so on [3].

MARS modeling is determined by trial and error for the combination of BF, MI, and MO to get the value of minimum GCV. GCV equation is as follows:

$$GCV(M) = \frac{ASR}{\left[1 - \frac{\tilde{C}(M)}{N} \right]^2} = \frac{\frac{1}{N} \sum_{i=1}^{N} \left[y_i - \hat{f}_M(x_i) \right]^2}{\left[1 - \frac{\tilde{C}(M)}{N} \right]^2} \tag{14}$$

In the case of additive modeling suggested to use a value of d = 2, based on the decline in expectation value of ASR [3]. While suggests conventional value $d = 4$ [1]. The smaller the value of d, the larger models which will produce with more functions of the base, and conversely the greater the value of d, the smaller models which will produce with fewer basis functions.

F. HIV/AIDS

Human Immunodeficiency Virus (HIV) is a virus that decrease the body's immune system so that the people affected by this virus will be susceptible to various infections and then causes *Acquired Immune Deficiency Syndrome* (AIDS). There are about 5 -10 million people living with HIV who do not yet show any symptoms but as a potential source of infection. AIDS is a disease that is very dangerous because it has a case fatality rate of 100% in five years, meaning that within 5 years after diagnosis of AIDS in upholding then all people will die [11]. Factors that affect the survival of people with HIV/AIDS are age, gender, education level, status employment, status marital, history ARV, absolute CD4 count, opportunistic infections, functional status, stage, and treatment compliance.

III. METHODOLOGY RESEARCH

A. Data Source

The data used in this research is secondary data on the medical records of HIV/AIDS patients in one hospital counted 100 data. Variables used in this research are:

Y : Survival Time
X_1 : Age
X_2 : Gender
X_3 : Education level
X_4 : Status of jobs
X_5 : Marital status
X_6 : History ARV

X_7 : absolute CD4 levels
X_8 : Opportunistic Infections
X_9 : Functional Status
X_{10} : Stadium
X_{11} : Compliance therapy

B. Method Analysis

a. Determine the survival data that will be used to eliminate the data censored.
b. Describing the characteristics of patients with HIV/AIDS
c. Predicting survival data distribution using the smallest of Anderson-Darling value
d. Determining the baseline hazard function
e. Estimating the survival function and cumulative hazard function
f. Using the Cox PH models to get Martingale residual,
g. Doing plotting data to know the Martingale residual predictor variables.
h. Modeling Cox PH with MARS approach through the following steps:
    1. Modeling with MARS combined Basis Function (22, 33, 44), Maximum Interaction (1, 2, 3), and the Minimum observation (0, 1, 2, 3)
    2. Getting the best model based on the value of the minimum GCV
    3. Modeling Cox Proportional Hazard with MARS approach
    4. Interpretation models
    5. Determine the level of interest for each of the significant variables in the model
i. Summing up the results of the analysis

IV. ANALYSIS AND DISCUSION

A. Descriptive Statistics

Before the description of the characteristics of patients with HIV/AIDS, then the description of the survival data were used.

TABLE 1. DESCRIPTIVE DATA SURVIVAL

N Total	n censored	n observation
100	51	49

Table 1 shows that of the 100 data obtained, there are 51 data classified in the data censored, where this data must be removed because it cannot be used in the survival analysis. It can be concluded that the survival data in this study there are as many as 49 data.

TABLE 2. DESCRIPTIVE PATIENTS HIV/AIDS

Variable	Characteristics	Number	Variable	Characteristics	Number
Age	Toddlers (0-5 years)	5	Absolute CD4 levels	>350	1
	Children (5-12 years)	1		200-350	5
	Adolescents (12-23 years)	2		<200	43
	Adults (>23 years)	41	Opportunistic Infections	< 2	17
Gender	Female	25		> 2	32
	Male	24	Functional Status	Normal	9
Education	Higher	11		Ambulatory	6
	Primary	21		lying	34
	None	17	Stadium	Stage I	12
Jobs	Working	31		Stage II	8
	Not Working	18		Stage III	15
Marital Status	Married	27		Stage IV	14
	Not Married	22	Compliance therapy	Comply	22
ARV	Ever	19		non-compliant	27

History	Never	30			

Table 2 shows the ingredients HIV/AIDS patients who experience even failure are aged above 23 years of age or older. With a CD4 count of less than 200.

B. Distribution Estimates

Estimation of the distribution is used to determine the distribution of survival data were used. The distribution function was used to estimate the survival function and cumulative hazard function. The distribution function is also used to determine the baseline hazard function which is used in the modeling.

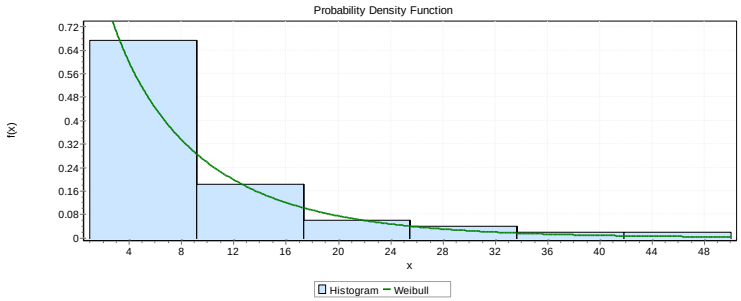


FIGURE 1. HISTOGRAM SURVIVAL DATA USED IN THE STUDY TO ESTIMATE THE DATA DISTRIBUTION.

Based on the estimation of the distribution using Anderson-Darling test, it is known that the smallest Anderson-Darling value is contained in a 2-parameter Weibull distribution in the amount of 1.93 with estimates of the parameters are and Based on estimates of parameters for two parameter Weibull distribution, then the baseline hazard function obtained are as follows :

$$\lambda_0(t|\eta,\gamma) = \frac{\gamma}{\eta}\left(\frac{t}{\eta}\right)^{\gamma-1}$$
$$= \frac{7.149}{0.859}\left(\frac{t}{0.859}\right)^{7.149-1}$$
$$= 8.322\left(\frac{t}{0.859}\right)^{6.149}$$

C. Estimated survival function and hazard function

Survival function is used to determine the probability of the patient's recovery, and cumulative hazard function is used to determine the rate of cure of HIV/AIDS. The estimation results of the survival function and the hazard function is as follows:

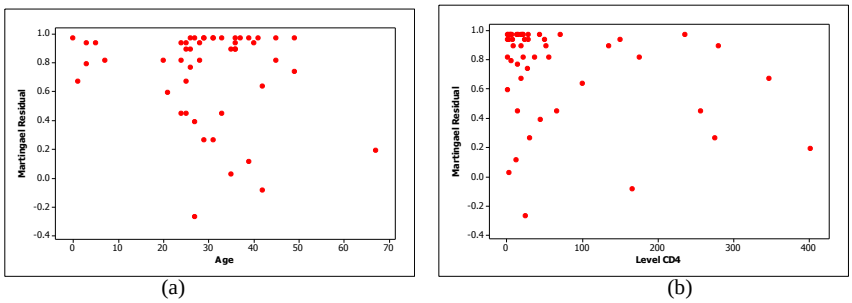
TABLE 3: ESTIMATED SURVIVAL FUNCTION AND CUMULATIVE HAZARD FUNCTION

Survival time	$S(t)$	$\Lambda(t)$	**Survival time**	$S(t)$	$\Lambda(t)$
1	0.969	0.031	14	0.666	0.406
2	0.939	0.063	15	0.575	0.553
3	0.899	0.106	16	0.542	0.612
5	0.831	0.185	18	0.478	0.738
7	0.811	0.209	24	0.445	0.810
8	0.791	0.234	27	0.412	0.887
11	0.769	0.262	28	0.377	0.976
12	0.720	0.328	36	0.337	1.087
13	0.694	0.365	50	0.281	1.269

Table 3 shows that the longer a patient is suffering from HIV/AIDS, the lower probabilities of survival for people is living with HIV/AIDS. On the contrary, the longer a patient suffering from HIV/AIDS, the higher the survival rate of patients with HIV/AIDS. It can be concluded that the probability of survival of patients with HIV/AIDS is inversely related to the survival rate of patients with HIV/AIDS.

D. Cox Proportional Hazard with Multivariate Adaptive Regression Spline

Before modeling with MARS, it is important to know the pattern of the relationship between the predictor variables and the response variable MARS modeling.



(a) (b)

FIGURE 2. SCATTER PLOT MARTINGALE RESIDUAL VS PREDICTOR VARIABLES (a) AGE, (b) THE ABSOLUTE CD4 LEVELS

Figure 2 shows that there is no particular pattern of variable X to variable Y. The pattern of relationships that do not tend to form patterns, showed that it could be used in MARS. MARS modeling done by trial and error for 36 combinations Basis Function (BF), Maximum Interaction (MI), and the Minimum Observation (MO) to get the best model based on the value of the minimum GCV.

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 = 0.573 with R2 = 0.729. Based on the results of this combination, it is known MARS models produced are as follows:

$$Y = 0.721 + 0.391 * BF3 + 0.200 * BF4 - 0.001 * BF5 - 0.015 * BF7 - 0.012 * BF10$$
$$+ 0112 * BF11 - 2.603 * BF12 + 0.017 * BF14 - 0.088 * BF16 - 0.010 * BF18;$$

Where:
BF2 = (X11 = 2);
BF3 = max (0, X1 - 35,000) * BF2;
BF4 = max (0, 35,000 - X1) * BF2;
BF5 = max (0, X7 - 1000) * BF3;
BF7 = max (0, 275,000 - X7) * BF2;
BF8 = (X2 = 1) * BF2;
BF10 = max (0, X7 - 25,000) * BF8;
BF11 = max (0, 25,000 - X7) * BF8;
BF12 = (X4 = 1) * BF8;
BF14 = max (0, X7 - 257 000) * BF4;
BF16 = (X3 = 3) * BF4;
BF18 = max (0, X7 - 236,000);

Resulting in a model hazard rate or the rate of survival of patients with HIV/AIDS as follows:

$$\lambda(t) = \lambda_0(t)\exp\left(\hat{Y}\right)$$
$$= 8.322\left(\frac{t}{0.859}\right)^{6.149} \cdot \exp\left(\begin{matrix}0.721 + 0.391 * BF3 + 0.200 * BF4 - 0.001 * BF5 - 0.015 * BF7 - 0.012 * BF10 \\ +0112 * BF11 - 2.603 * BF12 + 0.017 * BF14 - 0.088 * BF16 - 0.010 * BF18;\end{matrix}\right)$$

TABLE 4: INTERACTION ON BASIS FUNCTION

BF	Interactions	Specification
3 and 4	x1 and x11	Age and compliance
5	x7 * x1 and x11	CD4 levels, age and compliance
7	x7 and x11	CD4 levels and compliance
10	x2 and x11	Gender and compliance
11	x7 and x2	CD4 levels and gender
12	x7 * x2 and x11	CD4 levels, gender and adherence

14	x4 * x2 and x11	Employment, gender and adherence
16	x3* x1 and x11	The level of education, age and compliance

The Modeling results show that in general, the variables that affect the survival of patients with HIV/AIDS there are six variables: X1 (Age), X2 (Gender), X3 (Level of Education), X4 (Employment Status), X7 (Kadar CD4) and X11 (Compliance Therapy). The sixth of these variables has a good influence on the model, either individually or when interacting with other variables.

Table 4 shows the interaction of the variables that affect the survival of patients with HIV / AIDS. As for the variables that influence individual is adherence therapy and education level.

TABLE 5. VARIABLE INTEREST RATE

Variable	Importance	GCV
Gender	100	0.147
Therapy adherence	84.01	0.112
Employment Status	79.913	0.104
CD4 levels	78.947	0.102
Age	68.334	0.084
Education	16.149	0.032

Table 5 shows that gender have the largest contribution to the resulting model 100%. Then, the second largest contribution is in the amount of therapy adherence 84.010%. Then the third largest contribution is the employment status 79.913%, then the fourth biggest contribution is the Absolute CD4 cell count of 78.94%, the fifth biggest contribution is the Age of 68.334%, and the sixth biggest contribution is the level of education, amounting to 16 149.

V. CONCLUSION

HIV/AIDS patients who died is the average adult aged 23 years or older (age of majority), with CD4 levels below 200. Based on the modeling results with Cox Proportional Hazard MARS approach, which used a combination Basis Functions, Maximum interaction and minimum His observations are 22, 3, and 1 with a minimum GCV value was 0.028. Variables influencing the survival of patients with HIV/AIDS in individuals is age and compliance, levels of CD4 and compliance, gender and adherence, levels of CD4 and gender, CD4 count, gender and adherence, CD4 count, age and compliance, employment, gender and adherence, education level, age and compliance. Gender have the largest contribution to the resulting model, by 100%. Then, the second largest contribution is in the amount of therapy adherence 84.010%. Then the third largest contribution is the employment status of 79.913%, then the fourth biggest contribution is the Absolute CD4 cell count of 78.947%, the fifth biggest contribution is the Age of 68.334%, and the sixth biggest contribution is the level of education, amounting to 16.149%.

REFERENCES

[1] Kleinbaum. D. G. (2012). *Survival Analisis*, London, Springer

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

[3] Friedman, J.H., (1991), "Multivariate Adaptive Regression Spline", *The Annals of Statistics*, Vol. 19, pp 1-141.

[4] Kriner, M. (2007). Survival Analisis with Multivariate Adaptive Regression Splines. Disertasi. Munchen University.

[5] Nisa', F.S. dan Nudiantara (2012). Analisis Survival dengan Pendekatan Multivariate Adaptive Regression Spline pada Kasus Demam Berdarah Dengue (DBD). *Jurnal Sains dan Seni ITS.* Vol. 1, No. 1, 318-323

[6] Irwansyah, E. Nyoman, D.A, dan Bekti R.D. (2014). Cox Proportional Hazard with Multivariate Adaptive Regression Spline to Analyze the product Sales Time in E-Commerce. *Article in International Journal of Applied Mathematics and Statistics*

[7] Saputro. A. S. (2013) pemodelan *mixture survival* untuk kasus HIV/AIDS. Universitas Airlangga. Surabaya

[8] Collect, D. (2003). *Modeling Survival Data in Medical Research*. London: Chapman & Hall/CRC

[9] Purhadi. (2012). Analisis Survival Faktor-faktor yang mempengaruhi Laju kesembuhan pasien Penderita Demam Berdarah Dengue (DBD) di RSU Haji Surabaya dengan Regresi Cox. *Jurnal Sains dan Seni ITS*, Volume I. No. I., 271-267.

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

[11] Wibisono B, (1989). *Epidemiologi AIDS*; petunjuk untuk petugas kesehatan, Departemen Kesehatan RI. Jakarta.