Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 10, Number 4 (2014), pp. 591-600

http://www.ripublication.com

Cox Regression Model for Lifetime of a Newborn Baby (Neonates) in Dr. Saiful Anwar Malang Hospital

Ani Budi Astuti

Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia, Jl. Veteran Malang 65145 Indonesia
Email: ani_budi@ub.ac.id

ABSTRACT

Indonesia is one country that has a quite high birth rate and a high infant mortality rate. Various cases of newborn baby (neonates) death indicated caused by various factors such as low birth weight, low gestational age and low apgar scores. Cox regression is a method for modeling the relationship between a response variable in the form survival time with one or more predictor variables that are discrete or continuous. This paper proposes a model of the relationship between lifetime of newborn baby (neonates) with the predictor variables are birth weight, apgar scores and gestational age of neonates with Cox regression. The model is developed based on the data from a lifetime of newborn baby (neonates) were observed in DR. Saiful Anwar Malang Hospital. The results have succed to demonstrated Cox regression models for lifetime of a newborn baby (neonates) with the most suitable distribution is normal and has an average lifetime of 3.94406 days. Based on the test parameters model is known that a very close connection between lifetime of neonates with birth weight, apgar score and gestational age of neonates in which the positive association connection. Feasibility values of the Cox regression model were obtained AIC=655.936, $Q^2_{adjusted}$ =0.968 and Cox-Snell residual=6.888.

Keywords: Cox Regression, Model, Lifetime, Newborn Baby (Neonates).

1. INTRODUCTION

Indonesia is a developing country and has a quite high birth rate and a high infant mortality rate. Many cases of newborn baby (neonates) death indicated caused by various factors such as low birth weight, low gestational age and low apgar scores [1]. By knowing the linkage model of the relationship between lifetime of newborn baby

(neonates) with cause factors are expected to decrease the mortality rate of newborn baby in Indonesia.

Model for survival data is one of the generalized linear models. The model will establish the relationship between response variables in the form of the lifetime or the survival time from individuals in a certain condition [2] and [3]. Survival data models are widely applied in the modeling in the field of health [4]. Survival analysis is a technique analysis of survival data from one or several groups of individuals. Survival data is the data about the period of time between the timing of the beginning to the end time events. The duration of time between two events is defined as survival time and denoted by T [5] and [6].

Cox regression proposes a method for modeling the relationship between survival time of an individual with predictor variables were indicated to have a relationship with the survival time [7]. Predictor variable in the Cox regression method can be continuous or discrete (categorical). If it involves more than one predictor variables then have to qualify non-multicollinearity [8]. Cox regression model also known as the proportional hazards model because this model qualify the assumptions proportional hazard function. Proportional hazards is important assumption in the Cox regression i.e., the ratio between the two levels of hazard functions. Hazard function for level one is proportional to the hazard function for level two if the ratio of these two functions is constant and independent of time [9].

Research has been done previously associated with the data in this study is [1] and [10]. This paper proposes Cox regression models to determine the relationship between lifetime newborn baby (neonates) with predictor variables are birth weight, apgar scores and gestational age of the neonates were observed in DR. Saiful Anwar Malang Hospital.

2. MATERIALS AND METHODS

The data used in this study are data from 572 newborn babies in DR. Saiful Anwar Malang hospital as the data that have been used in the study [1]. The observed response variable is the lifetime of newborn baby (neonates) in the range of 0-7 days (Y). While the observed predictor variables associated with the response variable is birth weight neonates in grams (X_1), apgar scores in the range 1-8 (X_2) and gestational age with a range of 22-43 weeks (X_3). In addition, the observed variables status of neonates where a score of 1=dead and a score of 0=alive. General description of the observed data as in **Figure 1** to **Figure 4**.

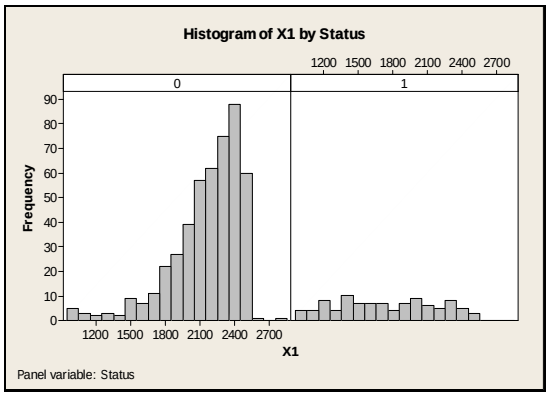


Figure 1. Histrogram of birth weight by status

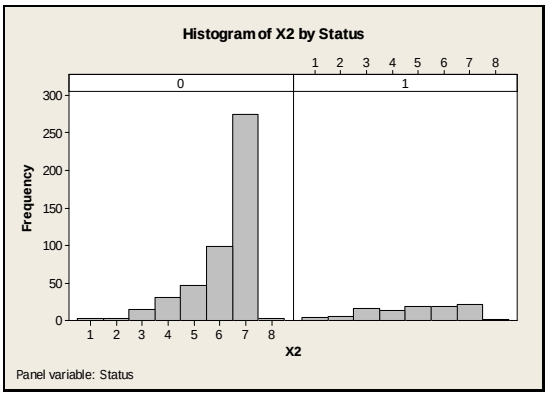


Figure 2. Histogram of apgar scores by status

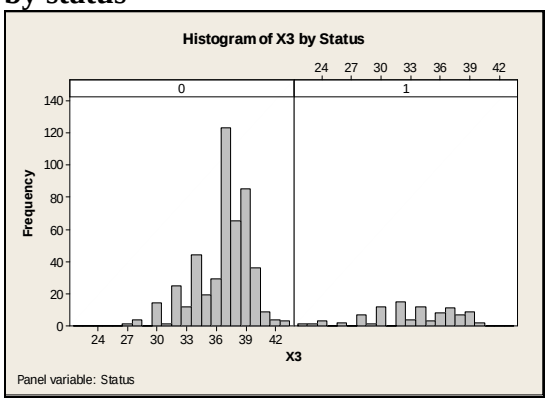


Figure 3. Histogram of gestational age by status

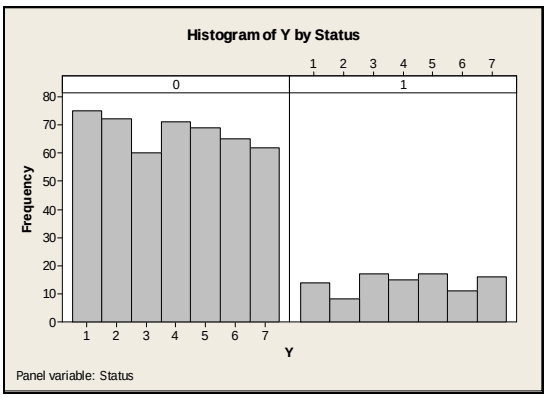


Figure 4. Histogram of neonates lifetime by status

2.1.Survival Function and Hazard Function

Survival function $S(t)$ is defined as the probability of an object has a survival time greater than *t*, *t* is defined as the actual survival time of an observed object and the value of the variable *T* has a non-negative value. It means that an object has a probability of living longer than *t* can be expressed as in **Equation (1)** [4]:

$$S(t) = P(T^3 t) = 1 - F(t). \quad (1)$$

$S(t)$ is a non-increasing function of time *t*. It is expressed as in **Equation (2)** [5], [6].

$$S(t) = \begin{cases} 1 & for\ t = 0 \\ 0 & for\ t = \infty \end{cases}. \quad (2)$$

Based on **Equation (2),** if $t = 0$ then $S(t) = S(0) = 1$. It means that none of the objects that have events as specified and the probability of survival an object will be worth one. If $t = \infty$ then $S(t) = S(\infty) = 0$, it means that if the test period increased until unlimited then in the end there will not be an object that can survive so the

probability of survival an object will approach the zero value. Graphically, the function $S(t)$ is illustrated in **Figure 5**.

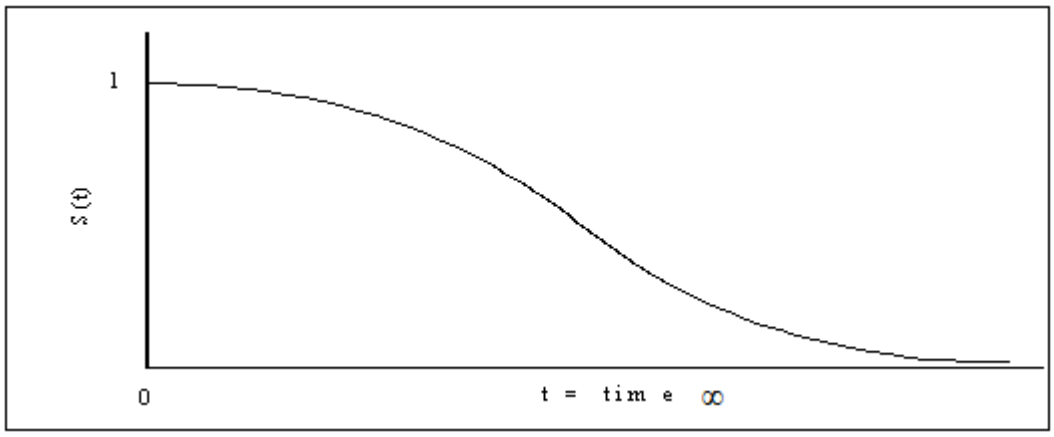


Figure 5. Plot of Survival Function $S(t)$ by *t*

Hazard function $h(t)$ is defined as the probability of an individual dies at time *t* on condition that the object has survived until the time *t*. This function is declared the rate of death of an individual has survived until the time *t*. For example, if the probability of survival time from an object is symbolized by T were among *t* and $t+\delta t$ with terms *T* greater than equal to *t*, written as $P(t \le T < t+\delta t \mid T \ge t)$. Conditional probablity is expressed as probability per unit of time divided by time interval δt as the state level. Hazard function $h(t)$ is the limiting value with δt approaching zero value. Hazard function can be expressed as in **Equation (3)** [4], [9], and [11].

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \le T < t+\delta t \mid T \ge t)}{\delta t} \right\} \tag{3}$$

If the survival function associated with the hazard function would be more meaningful if the **Equation (3)** converted into the **Equation (4)**.

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \ge T) - P(T > t+\delta t)}{P(T \ge t)\,\delta t} = \frac{P(t \le T < t+\delta t)}{P(T \ge t)\,\delta t} = \frac{F(t+\delta t) - F(t)}{S(t)\,\delta t} \tag{4}$$

2.2. Cox Regression Models

Cox proportional hazard model is commonly called the Cox regression has an important role in the survival analysis. The basis of Cox regression model is generated from the hazard function for the i^{th} object at the time *t* that consists of two factors, the baseline hazard function with symbolized as $h_o(t)$ and a linear function of a set of *k* predictor variables that were generated by the exponent. Generally, Cox regression model is defined as in **Equation (5)** [9].

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + ... + \beta_k x_{ik}), \quad (5)$$

where $h_o(t)$ function can be regarded as a hazard function for an object. If the predictor variables in **Equation (5)** is 0, then $\exp(\beta_1 x_{i1} + ... + \beta_k x_{ik})$ can written as $\exp(\eta_i)$ and referred to as the relative hazard where η_i referred to as linear combinations of predictor variables in the x_j with $j = 1, 2, ..., p$. Generally, model of Cox regression can also be written as in **Equation (6)**.

$$\left(\frac{h_i(t)}{h_0(t)}\right) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}), \quad (6)$$

where $h_0(t) = \frac{f(t)}{S(t)}$. $h_i(t)$ is the probability failure or death of i^{th} object at time *t*, $h_o(t)$ is baseline hazard function, $f(t)$ is the density function of resistance probability to object-*t*, whereas $S(t)$ is survival function. Value of hazard function, $h_i(t)$ is determined after the $h_0(t)$ is obtained. Cox regression model showed that the mortality ratio between objects in the group indicated by $\exp(\beta_j)$ multiplied by the ratio of deaths between objects in the control group continuously [11]. Method of parameters estimation in the Cox regression model can be estimated by using the maximum partial likelihood method [4] and [7]. Based on Newton-Raphson procedure, estimation of parameters β on $s+1$ is symbolized $\hat{\beta}_{s+1}$ as in **Equation (7)**.

$$\hat{\beta}_{s+1} = \hat{\beta}_s + \mathbf{I}^{-1}(\hat{\beta}_s)u(\beta_s), \quad (7)$$

where $\mathbf{I(\beta)} = -\frac{\partial^2 \log L(\beta)}{\partial\beta_j \partial\beta_k}$, (8)

with $j = 1, 2, ..., p$ and $k = 1, 2, ..., p$. **I(β)** is called the Hessian matrix or information matrix observations.

The testing for the significance of the parameters of Cox regression model include simultaneous and partial test. Simultaneous test is used to examine the influence together of predictor variables on the response variable. The test uses the likelihood ratio test with the hypothesis test is as follows [5] and [6]:

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ versus H_1 : at least there is one $\beta_j \neq 0$, for $j = 1, 2, ..., p$.

The test statistics is as in **Equation (9)**.

$$\chi^2_{LR} = 2\left[Ln\, L(\hat{\beta}) - Ln\, L(\beta_0) \right]. \quad (9)$$

If $\chi^2_{LR} > \chi^2_{p,\alpha}$ then H_0 is rejected.

Partially test is used to examine the effect of each predictor variable on the response variables. This test is done by dividing the parameter estimator with the standard error of the parameter estimator. These ratio is called the Wald statistic and the hypothesis test of Wald test is as follows [5] and [6]:

$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, for $j = 1, 2, ..., p$.

The Wald test statistics is as in **Equation (10)**.

$$W = \frac{\hat{\beta}_j^{\,2}}{\left[Se(\hat{\beta}_j) \right]^2} \sim \chi_1^2, \tag{10}$$

where

$\hat{\beta}_j$: regression coefficient on the variable j^{th}

$\left[Se(\hat{\beta}_j) \right]^2$: varians of regression coefficients.

If $W > \chi_1^2$ then H_0 is rejected.

Identity of residual on Cox regression model aimed to the suitability of the model. Cox-Snell residual is often used in testing the Cox regression model. Cox-Snell residual can be interpreted as the expected value of each observation. Cox-Snell residual for the i^{th} individual can be formulated as in **Equation (11)** [12].

$$r_{Ci} = \exp\left(\hat{\beta}' x_i \right) \hat{H}_0 \left(t_i \right), \tag{11}$$

where

r_{Ci} : Cox-Snell residual for the i^{th} individual

$\hat{H}_0 \left(t_i \right)$: estimate of cumulative baseline hazard function at time t_i.

The suitability of the model can also use Akaike's Information Criterion (AIC) developed by Hirotsugu Akaike in 1971. AIC value is a measure of the goodness estimators statistical model taking into account the number of parameters in the model [12]. Generally, AIC can be formulated as in **Equation (12)**.

$$AIC = 2k - 2\ln(L), \tag{12}$$

where:

k : number of parameters in the model

L : maximum likelihood model that allegedly.

$Q^2_{adjusted}$ value is used to measure how accurate the predictions of the model formed by considering the number of predictor variables in the model. $Q^2_{adjusted}$ value can be obtained by the formula as in **Equaion (13)** [12].

$$Q^2_{adjusted} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 \,/\, n-k}{\sum_i (y_i - \bar{y})^2 \,/\, n}, \tag{13}$$

where:

$Q^2_{adjusted}$: the adjusted value of cross validation

k : many predictor variables that form

y_i : variable response of the i^{th} observation

$\hat{y}_i$: the predicted value for y_i

$\bar{y}_{.}$ average response variables

$Q^2_{adjusted}$ scale value to the range $0 < Q^2_{adjusted} < 1$, where $Q^2_{adjusted}$ is getting close to 1 means that the model obtained yield more accurate predictions.

3. RESULT AND DISCUSSION

3.1.Test Assumptions to Non-multicollinearity

Multicollinearity test results to predictor variables by Variance Inflation Factor (VIF) is as in **Table 1**.

Table 1. Result of Multicollinearity Test to Predictor Variables

Predictor	VIF	Result of Test
X_1	1.374	Non-multicollinearity
X_2	1.137	Non-multicollinearity
X_3	1.274	Non-multicollinearity

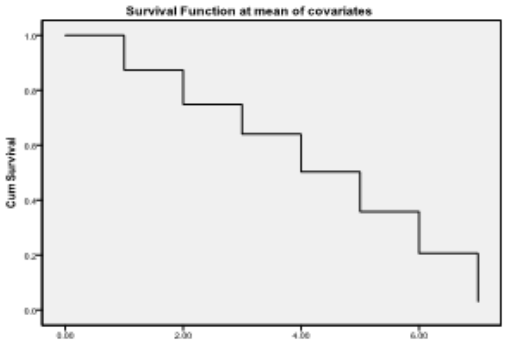
Based on **Table 1,** it can be seen that all predictor variables are independent because VIF values <5.

3.2.Identification of Neonates Lifetime Distribution

Identification of Neonates lifetime distribution is very important because the selection of appropriate methods of survival analysis can be done through distribution of lifetime neonates variable. The test results can be seen in the **Table 2**.

Table 2. Result of Suitability Test of Neonates Lifetime Distribution

Goodness-of-Fit Distribution	Anderson-Darling (adj)	Correlation Coefficient
Weibull	17.697	0.940
Lognormal	28.521	0.933
Exponential	130.280	*
Loglogistic	33.471	0.914
3-Parameter Weibull	18.004	0.940
3-Parameter Lognormal	15.419	0.960
2-Parameter Exponential	91.037	*
3-Parameter Loglogistic	22.970	0.937
Smallest Extreme Value	39.477	0.918
Normal	15.413	0.960 with MTTF value at 3.94406
Logistic	22.962	0.937



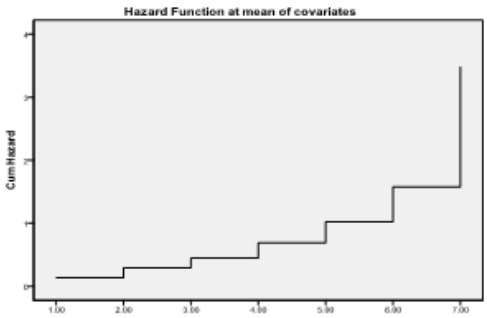


Figure 6. Survival function of neonates lifetime (0-7 days)**Figure 7.** Hazard function of neonates lifetime (0-7 days)

Based on **Table 2,** it can be seen that the normal distribution is the most suitable distribution for lifetime of neonates because it has the smallest value of the goodness of fit at 15.413 and the average lifetime of neonates at 3.94406 days. Information of **Figure 6** shows that the longer the lifetime of neonates then declined the probablity of his life. Meanwhile, according to **Figure 7**, the rate of death increased when the lifetime of neonates increased.

3.3.Parameter Estimation and Significance Test of Cox Regression Model

Results of parameter estimation and testing of parameters from the Cox regression model with normal distribution approach can be seen in **Table 3**.

Table 3. The Results of Parameter Estimation and Test of Cox Regression Model

Regression with Life Data: Y versus X_1, X_2, X_3

```
Response Variable: Y
Censoring Information  Count
Uncensored value          98
Right censored value     474
Censoring value: Status = 0
Estimation Method: Maximum Likelihood
Distribution:   Normal
Regression Table
                      Standard                    95.0% Normal CI
Predictor       Coef      Error      Z      P      Lower      Upper
Intercept   -5.39951    1.82204   -2.96  0.003   -8.97064   -1.82838
X1          0.0013452  0.0005514   2.44  0.015    0.0002644  0.0024260
X2          0.706857   0.121809    5.80  0.000    0.468115   0.945598
X3          0.181053   0.0568938   3.18  0.001    0.0695433  0.292563
Scale       2.66650    0.194071                   2.31202    3.07534
Log-Likelihood = -323.968
Anderson-Darling (adjusted) Goodness-of-Fit
Cox-Snell Residuals = 6.888
```

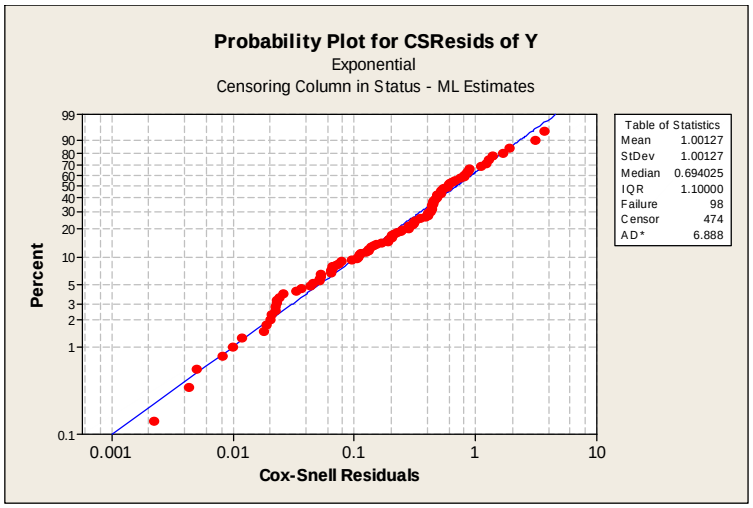


Figure 8. The plot of the Cox-Snell residuals Probability to Y

Based on **Table 3,** it can be seen that the estimators for the model parameters b_0,b_1,b_2,b_3 is -5.39951, 0.0013452, 0.706857, 0.181053 respectively. Whilst the results of significance test show that all parameters of the model have very significant with p value<5%. It means that all predictor variables in the model are very significant effect on survival time in which the greater the birth weight, apgar scores and gestational age of neonates, the greater the lifetime of neonates and conversely. The Cox regression model is:

$$h_i(t) = \text{-}5.39951\exp\left(0.0013452X_{i1} + 0.706857X_{i2} + 0.181053X_{i3}\right).$$

The model has the Cox-Snell residual value at 6.888 (small enough) and has the plot of Cox-Snell residuals probability to Y fit to the normal line (**Figure 8**). The AIC value is 655.936 and the $Q^2_{adjusted}$ value is 0.968. It can be said that the Cox regression model is a very good model in the describing relationship between lifetime of neonates with the predictor variables are birth weight, apgar scores and gestational age.

4. CONCLUSION

Cox regression model for neonates lifetime of the predictor variables birth weight (X_1), apgar scores (X_2) and gestational age (X_3) is $h_i(t) = \text{-}5.39951\exp\left(0.0013452X_{i1} + 0.706857X_{i2} + 0.181053X_{i3}\right)$ where all model parameters are very significant effect on lifetime neonates. The greater the birth weight, apgar scores and gestational age than the longer the lifetime of neonates and conversely the lower the birth weight, apgar scores and gestational age then the shorter the lifetime of neonates. The Cox regression model has a Cox-Snell residual value at 6.888, the AIC value at 655.936 and the $Q^2_{adjusted}$ value at 0.968.

5. ACKNOWLEDGEMENTS
We would like to thank to management of DR. Saiful Anwar Malang hospital which has provided the data and to anonymous refewer to this paper.

6. REFERENCES

[1] Niasari, M., 2002, "Birth Weight, Gestational Age and Apgar Scores as Predictors of Premature Death Neotal in Low Birth Weight Baby (Retrospective Cohort Study in RSUD Dr. Saiful Anwar Malang in 2000)." (in Indonesia). Unpublished Thesis in Partial Fulfillment of the Requirements for the Degree of Bachelor of Medicine, University of Brawijaya, Malang.
[2] McCullagh, P. and Nelder, J. A., 1997, Generalized Linear Models, Second Edition, Chapman & Hall, London.
[3] Baig, M. A. K., Dar, J. G., and Mir, M. I., 2009, "Generalized Past Entropy in Survival Analysis," Global J. Of Pure and Applied Mathematics, 5(3), pp. 201-208.
[4] Collet, D., 2003, Modeling Survival Data in Medical Research Second Edition, Chapman and Hall, London.
[5] Miller, R. G., 1998, Survival Analysis, John Willey and Sons, New York.
[6] Lee, E. T., 1997, Statistical Methods for Survival Data Analysis, Belmont, CA: Wadsworth.
[7] Cox, D. R., 1972, "Regression Model and Life Table (With Discussion)," Journal of the Royal Statistical Society, B, 74, pp. 187-220.
[8] Kleinbaum, D. G., and Klein, M., 2005, Survival Analysis: a Self-Learning Text, Second Edition, Springer Verlag, New York.
[9] Fox, J., 2002, "Cox Proportional-Hazard Regression for Survival Data." Appendix to An R and S-PLUS Companion to Applied Regression.
[10] Wafah, A. Z., 2006, "Classification Tree Method on Binary Response Data," (in Indonesia), Unpublished Thesis in Partial Fulfillment of the Requirements for the Degree of Bachelor of Statistics, University of Brawijaya, Malang.
[11] Chan, Y. H., 2004, "Biostatistics 203: Survival Analysis," Singapore Med J, 45 (6), pp. 249-256.
[12] Anderson, P. K., 1982, "Testing Goodness of Fit of Cox Regression and Life Model," Biometrics, 38, pp. 67-77.