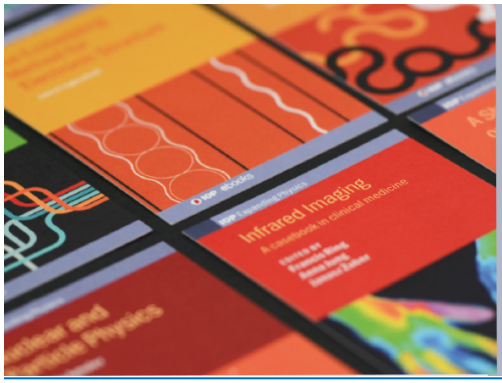
Journal of Physics: Conference Series

Cox Proportional Hazard Regression Analysis of Dengue Hemorrhagic Fever



IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012242 doi:10.1088/1742-6596/1028/1/012242

Cox Proportional Hazard Regression Analysis of Dengue Hemorrhagic Fever

Suwardi Annas[1], M. Nusrang[2], R. Arisandi[3], N Fadillah[4], and P Kartikasari[5]

[1,2,3]Department of Statistics, Universitas Negeri Makassar,Makassar,Sulawesi Selatan,90222,Indonesia
[4,5]Departement of Information System, Stikom Surabaya,Surabaya,Jawa Timur,60298,Indonesia

suwardi_annas@unm.ac.id

Abstract: Survival analysis is a statistical procedure used to analyze data in the form of time of the incident until an event occurs. The purpose of the survival analysis is to estimate the probability of survival, recurrence, death or other events until a certain period of time. .In this case, one of the regression method that can be used Cox Proportional Hazard (Cox PH) regression. Data used in this research is data of 70 patients Dengue Hemorrhagic Fever (DHF) in Makassar City Hospital. The results of analysis indicated that the factor that most affect the rate of healing DHF patients is platelet factor. The rate of recovery of patients with DHF with platelet counts below normal is 2,625 times the normal platelet count. Therefore, patients with dengue disease who have lower platelet counts tend to have a longer recovery rate than patients who have normal platelet counts.

1. Introduction

Survival analysis is a statistical procedure used to analyze data where the variables are the time to the occurrence of an event [9]. Time is stated from the beginning to observation until an event occurs [2]. The purpose of this analysis is to estimate the probability of survival, recurrence, death or other events until a certain period of time and to determine the relationship between the time of occurrence which is the dependent variable and the independent variables measured at the time of the study.[6,8]. In addition, it can also be used to see which factors are most influential on an event or event [5,13].

Classical regression analysis can be used to study relationships between dependent and independent variables. However, this classical regression analysis requires the assumptions of distribution and form of functional relationships among known variables [3,7]. In practice empirically this assumption is not always appropriate because it is possible, the data obtained can not provide information on the distribution of survival time, so that the form of functional relationship of the basic hazard function also can not be known.

To solve the problem in this research will be used Cox Proportional Hazard (Cox PH) regression method which does not depend on distribution assumption from time of its incident. In addition the Cox PH model is semiparametric distributed so it does not require information about the underlying distribution of survival time and regression parameters can be estimated from the model. The semiparametric model can also be used although the functional form is unknown, since the Cox PH

model can still provide hazard ratio (HR) information that is independent of its functional relationship defined as the ratio of an individual hazard rate (HR) to the hazard rate of another individual [10].

The results of the Cox PH model are almost the same as the parametric model results and can estimate the hazard ratio. In addition, this model is a safe model chosen when in doubt to determine its parametric model, so there is no doubt about the choice of the wrong parametric model[9]. This regression is more popularly used in research on health data which is bound in time. For example data about the time of the patient suffering from a particular disease, which starts from the beginning of the hospital entrance to certain events, such as death, heals or other special occurrence. In this study the application of this method is used in patients with Dengue Hemorrhagic Fever (DBD) in Makassar City to determine the factors that affect the rate of recovery of patients..

2. Material and Method

The data used in this research is the data of DHF patients identification of the variables set as the criteria of DHF patients obtained from the medical records of the Regional General Hospital (RSUD) Makassar City in 2015. The variable used is the length of inpatients as a variable Y, while age (X1) , gender (X2), hemoglobin (X3), leukocytes (X4), hematocrit (X5), platelets (X6), and body temperature (X7). The stages of research can be described as follows.

2.1. Test the assumption of data distribution and PH.

Before to estimate the parameters of Cox PH regression models first test the distribution and assumptions PH. The distribution of survival time should be detected to form Hazard function, while testing the data distribution is done by Kolmogorov Smirnov test with test statistic D= sup

$$|S(x) - F_0(x)|$$

While checking the assumption PH with global test test.

2.2. Estimation and Test of Parameter significance

The coefficient of $\boldsymbol{\beta}$ in the Cox PH model is estimated using the Maximum Likelihood Estimator (MLE) method. With likelihood function

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta x_{(i)})}{\sum_{i \in R(t_i)} \exp(\beta x_{(i)})} \tag{2.1}$$

$x_{(i)}$ is a variable vector of individuals failing at time-i with time t_i., and $R(t_i)$ = all individuals who are at risk of failure at a time-i.

After the likelihood function is formed, it then makes the ln-likelihood function,

$$\ln L(\beta) = \ln\left(\prod_{i=1}^{r} \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{j(i)}\right)}{\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)}\right) = \sum_{i=1}^{k} \left[\left(\sum_{j=1}^{p} \beta_j x_{j(i)}\right) - \left(\ln \sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)\right)\right] \tag{2.2}$$

then look for the first derivative to be equal to zero

$$\sum_{i=1}^{k} \left[\left(\sum_{j=1}^{p} x_{j(i)}\right) - \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)}{\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)}\right] = 0 \tag{2.3}$$

and look for the second derivative.

$$= \sum_{i=1}^{k} \left[\frac{\left(\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right) \sum_{j=1}^{p} x_{jl}\right)^2}{\left(\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)\right)^2} - \frac{\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)\left(\sum_{j=1}^{p} x_{jl}\right)^2}{\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)}\right] \tag{2.4}$$

Since the equation is not close form then to get the estimator value is done by numerical method that is with Newton Raphson iteration, then the parameter estimation on iteration to (c + 1) following.

$$\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{c+1}}\right)_{px1} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{I^{-1}}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}}\right)_{pxp} \boldsymbol{U}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}}\right)_{px1} \tag{2.5}$$

Iteration will stop if, $\left\|\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{c+1}}\right) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{c}}\right\| \leq \in$, where $\in$ is a very small number (convergent).

There are two ways to test the significance of parameters, including concurrent tests and partial tests. For simultaneous test with test statistic

$$G = -2[\ln L(0) - \ln L(\hat{\beta}_k)] \tag{2.6}$$

.

As for the Partial test using the test statistic

$$W = \left|\frac{\hat{\beta}_k}{SE\ (\hat{\beta}_k)}\right|^2 \tag{2.7}$$

with $\hat{\beta}_k$: coefficient of parameter estimator and $(\hat{\beta}_k)$: standard error parameter estimator $\hat{\beta}$ [12].

2.3. Best Model Selection

In the best model selection, this study uses Akaike Information Criterion (AIC). The best model has the smallest AIC value

$$AIC = -2\log\hat{L} + 2P \tag{2.8}$$

2.4. Hazard Ratio

The rate of recovery of patients can be seen from the value of hazard ratio or odds ratio. aLet X be an independent variable with two categories. i.e. 0 and 1. According to [1], the relationship between the variable X with h (t) is expressed by $h_0\ (t|x) = h_0(t)e^{\widehat{\beta x}}$, then:
Individual with x = 1, hazard function:

$$h_0\ (t|x = 1) = h_0(t)e^{\widehat{\beta.1}} = h_0(t)e^{\hat{\beta}} \tag{2.9}$$

Individual with x = 0, hazard function:

$$h_0\ (t|x = 0) = h_0(t)e^{\widehat{\beta.0}} = h_0(t) \tag{2.10}$$

Hazard ratio for individuals with x = 0 compared x = 1 is :

$$Hazard\ Ratio = \frac{h_0\ (t|x=0)}{h_0\ (t|x=1)} = \frac{h_0(t)}{h_0(t)e^{\hat{\beta}}} = e^{\widehat{-\beta}} \tag{2.11}$$

3. Result and Discussion

3.1. Test of Distribution and PH Assumption

Test of distribution assumption in table 1 for patient's length of stay, it is found that p-value is greater than α = 0,05. This result shows that the data does not follow the distribution, either weibull, lognormal, or exponential. Since the data do not follow the distribution, we can proceed with the semiparametric model using the Cox regression.

Table 1 Test of Distribution Assumption in Survival Time Data

Distribution	p-value
Weibull	0,009
Lognormal	0,001
Exponential	0,001

Further testing of Proportional Hazard (PH) assumptions in this study using global test test. In table 2, it shows the global test results for each variable. Each variable has a p-value > 0,05, it can be concluded that the assumption of PH is met for all independent variables.

Because in this case it does not follow either the parametric model whether it is Weibull or exponential, but fulfill the proportional assumption of Hazard (Cox PH) then the analysis used is Cox regression analysis of PH which is a semi parametric model and does not require information about the underlying distribution of survival time so that regression parameters can be estimated from the model [9,10].

Table 2 The value of p-value Global Test

Variable	p-value
Age (X1)	0,0638
Gender (X2)	0,2414
Hemoglobin (X3)	0,7105
Leukocytes(X4)	0,6601
Hematocrit (X5)	0,6803
Platelets (X6)	0,5848
Body Temperature (X7)	0,6451
GLOBAL	0,5026

3.2. Parameter Estimation and Parameter Testing

One of the objectives of the Cox PH model is to model the relationship between survival times and the variables suspected to affect survival time. The parameter parameters in the Cox PH model can use the Maximum Likelihood Estimator (MLE) method. The test parameters include simultaneous testing and partial testing. The result of simultaneous test shows that the value of Likelihood partial test is 17,27, with P-value equal to $0{,}015 < \alpha = 0{,}05$. These results conclude that at least one of the significant variables, or can be interpreted that the overall model can contribute to the rate of recovery of patients with DHF.

Furthermore, partial testing can be seen in table 3. The result of parameter test shows that the variable has p-value < $\alpha = 0{,}05$ only platelet variables, so it can be concluded that platelet variables are significant for the rate of recovery so that it can contribute to the healing rate of patients with DHF. While other variables have no significant effect.

Table 3 Partial Test Results for All Modifiers

Variable	$\hat{\beta}$	p-value	Information
Age (X1)	0,006	0,572	not significant
Gender (X2)	-0,152	0,470	not significant
Hemoglobin (X3)	-0,251	0,098	not significant
Leukocytes(X4)	0,042	0,206	not significant
Hematocrit (X5)	0,094	0,110	not significant
Platelets (X6)	-0,805	0,030	Significant
Body Temperature (X7)	-0,175	0,066	not significant

3.3. Selection of the Best Model

After making parameter estimation and variable significance test, then selected the best model for the resulting model has a minimum error. In this study the method used is Backward Selection with the process of elimination of variables entered into the model, starting with removing or deleting one by one according to the criterion of significance. The best model is seen from the smallest AIC value of the model formed. Based on table 4 shows that the best model formed is a model with two independent variables i.e. thrombocyte variable (X6) and body temperature (X7) with the lowest AIC value of 747,06.

Table 4. AIC Value for Best Model Selection

Model	Variable	AIC
1	All free variables are included	752,08
2	No age variables (X1)	750,38
3	Without age (X1) and gender (X2)	748,89
4	Without age (X1), sex (X2), and leukocyte (X4)	748,42
5	Without age (X1), sex (X2), leukocyte (X4), and hematocrit (X5)	748,06
6	Without age (X1), sex (X2), leukocyte (X4), hematocrit (X5), and hemoglobin (X3)	747,06

After the selection of the best model formed is a model with two independent variables of the platelet variables (X6) and body temperature (X7). From the results of parameter estimation in table 5, the best Cox PH regression model used is:

$$h(t) = h_0(t) \exp(-0{,}96418\, X_6 - 0{,}17331\, X_7) \quad (3.1)$$

Table 5 Cox Regression Parameter PH of the Best Model

Variable	$\hat{\beta}$	p-value
Platelets (X6)	-0,9642	0,0061
Body Temperature (X7)	-0,1733	0,0539

3.4. Hazard Ratio

The rate of recovery of patients with DHF disease can be determined by finding the value of hazard ratio or odds ratio of the variables included in the best model. Hazard ratio is one of the important parameters in survival analysis [4]. The value of the hazard ratio is a measure used to determine the level of risk (trend) that can be seen from the comparison between individuals with the condition of free X variables in the category of success with the category failed. Let X be an independent variable with two categories. i.e. 0 and 1.

Based on table 6, the platelet variables in the normal and under normal category. The parameter estimate for platelet hazard ratio was 2.625. This hazard ratio value means that the rate of recovery of patients with DHF with platelet counts below normal is 2,625 times the normal platelet count. Therefore, patients with dengue disease who have lower platelet counts tend to have a longer recovery rate than patients who have normal platelet counts. These results are in line with other research conducted by [5], the results obtained there is a significant relationship between platelet count and patient recovery rate. As for the temperature of the body, because the value of hazard ratio is worth 1.189, which means more than 1, the higher the body temperature of a patient, then the patient's recovery rate is longer

Table 6 Hazard Ratio for Platelet and Body Temperature

Variable	$\hat{\beta}$	Hazard Ratio ($e^{-\hat{\beta}}$)
Platelets (X6)	-0,9642	2,625
Body Temperature (X7)	-0,1733	1,189

Based on the results of research in table 3.6 then modeling regression Cox PH in cases of Dengue Hemorrhagic Fever (DHF), it can be concluded as follows:
Cox PH regression modeling for patient data of dengue fever patients at RSUD Kota Makassar in 2015 is as follows.

$$h(t) = h_0(t) \exp(-0{,}9642\ trombosit - 0{,}1733\ suhu\ badan) \quad (3.2)$$

4. Conclusion

This study has applied Cox Proportional Hazard analysis in the case of DHF patients where the data distribution does not follow one of the distribution of paremetric data. One of the advantages of this method if it is bathed by a classical regression analysis is because the Cox regression of Proportional Hazard does not depend on the assumption of the distribution of the timing of an event. In addition the Cox PH model is semiparametric distributed and therefore does not require information about the underlying distribution of survival time.
In this study, factors that allegedly contribute to the occurrence of dengue virus-induced diseases include age, gender, hemaglobin count, leukocyte, hematocrit, platelets, and body temperature. The results of this study conclude that the variables that most affect the rate of healing of patients with DHF in RSUD Kota Makassar in 2015 is the number of platelets of patients, while other variables have no significant effect.

References

[1] Aly EAA, Kochar SC, and McKeague IW (1994). Some tests for comparing cumulative incidence functions and cause-speci¯c hazard rates. JASA 89, 994-999.
[2] Collectt, D. (2003). Modelling Survival Data In Medical Research "second edition". Chapman & Hall: New York.
[3] Draper, N., & Smith, H. (1992). Analisa Regresi Terapan, Second Edition. New York: John Wiley & Sons
[4] Dahlan, M. S. (2013). Analisis Survival "Dasar-Dasar Teori dan Aplikasi Program Stata". Sagung Seto: Jakarta.
[5] Ernawatiningsih, N. P. L. (2012). Analisis Survival Dengan Model Regresi Cox. Jurnal Matematika, 2 (2), 1693-1394.
[6] Hosmer, D. W., Lemeshow, S., & Mya, S. (2008). Applied Survival Analysis:Regression Modelling of Time to Event Data. New Jersey: John Wiley
[7] Hocking, R. (2003). Methods and Application of Linear Models (Regression and The Analysis of The Variance, Second Edition). New York: John Wiley & Sons.
[8] Jenkins, S. P. (2005). Survival Analysis. Unpublished Manuscrip: New York.
[9] Kleinbaum, D. G., & Klein, M. (2005). Survival analysis : a self learning text. Springer-Verlag: New York.
[10] Lee, E. T., & Wang, J. (2003). Statistical methods for survival data analysis. John Wiley & Sons: Canada.
[11] Law, A. M., & Kelton, D. W. (2000). Simulation Modelling Analysis ($3^{th\ ed}$.). New York: MacGraw-Hill
[12] Mohammed, D. M. A. (2014). Survival Analysis By Using Cox Regression Model With Application. International journal of scientific & technology research, 3(11), 277-8616..
[13] Tustianto, Kris., & Soehono, L. A. (2012). Pemodelan regresi Cox proportional hazard faktorfaktor lama proses IMB Kota Malang. Jurnal matematika.