A matlab code to compute prediction of survival trends in patients with DHF

IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012113 doi:10.1088/1742-6596/1028/1/012113

A matlab code to compute prediction of survival trends in patients with DHF

B Poerwanto[1*], R Y Fa'rifah[1], W Sanusi[2], S Side[2]

[1]Department of Informatics Engineering, Univeristas Cokroaminoto Palopo, Palopo 91913, Indonesia
[2]Department of Mathematics, Universitas Negeri Makassar, Makassar 90222, Indonesia

*bobbypoerwanto@uncp.ac.id

Abstract. This study aims to create matlab code in estimating the parameters of cox regression model. The written matlab code consists of two algorithms. the first algorithm is to obtain the parameter estimation at the baseline hazard function derived from the weibull distribution while the second is on the cox model.The parameter estimation used in the first algorithm uses maximum likelihood estimation (MLE) and the second algorithm uses the maximum partial likelihood estimation (MPLE). The written matlab code is simulated to predict the survival time of DHF patients. The results is that age and thrombocyte are significant covariates affecting survival time of patients with DHF.

1. Introduction

Cox regression is a semiparametric model [1],[2] used to determine the relationship between covariate variables and survival data [3],[4]. This regression is often used in clinical studies to analyze survival time, as did by Omurlu et al [2],[4] in cases of breast cancer, Ahmed et al [1] in the case of colon cancer, and Side et al [5],[6],[7] in the case of hepatitis and TB. The cox regression model consists of two functions, namely the baseline hazard function and the exponent function, ie the functions that contain the coefficient of the cox regression. The baseline hazard function is a function of a certain distributed survival time, such as weibull, exponent, and lognormal.

There are several methods for estimating parameters of cox regression. Devarajan and Ebrahimi [8] using two methods: B-Splines cubic to estimate baseline hazard and maximum likelihood estimation to obtain parameter coefficients on non-proportional hazard. Meanwhile, Gradowska and Cooke [9] estimated the parameters of the model using least square method in cox proportional hazard. The method is used to estimate the parameters of the overall cox regression model on unencensored data. So, the model formed is like a linear regression model. Other researchers such as Ojeda et al [10], Omurlu et al [4], and Chen et al [11] used a maximum partial likelihood estimation (MPLE) which is estimation method without involving baseline hazard function

Based on Ojeda et al [10], Omurlu et al [4], and Chen et al [11], this study used the MPLE method to obtain the regression coefficients of the cox regression model. The baseline hazard function in this study is a hazard function of survival time which has weibull distribution, in contrast to Devarajan and Ebrahimi [8]. In this study, the parameters in the baseline hazard function are obtained from estimating survival time parameters by using MLE.

The parameter estimation of the two functions is written on the matlab code consisting of two algorithms. The first algorithm is to obtain parameter estimation in the baseline hazard function while the second algorithm is on the regression coefficient. The algorithms are applied to predict hazard rate of DHF patients with covariate used are age, sex, hemoglobin, leucocytes, hematocrit and thrombocyte.

2. The Cox Regression Model

Let T be the variable describing the survival time of n sample. $T=\left(t_1,t_2,\ldots,t_n\right)^T, T>0$, $\mathbf{X}$ is p covariate where $\mathbf{X}=\left(x_{i1},x_{i2},\ldots,x_{nj}\right)^T, i=1,2,\ldots,n\ j=1,2,\ldots,p$. The cox regression model can be written as:

$$h\left(t|X\right)=h_0\left(t\right)\exp\left(\boldsymbol{\beta}^T\mathbf{X}\right) \tag{1}$$

where $h_0\left(t\right)$ is a baseline hazard function which is the function following a certain distribution, and $\boldsymbol{\beta}$ is the coefficient of size model $p\times 1, h_0\left(t\right)$ is obtained from the probability density function divided by survival function $S\left(t\right)$. If $h_0\left(t\right)$ follows weibull distribution, then:

$$f\left(t\right)=\frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}\exp\left(-\left(\frac{t}{\gamma}\right)^{\alpha}\right)$$

$$S\left(t\right)=1-F\left(t\right)=\exp\left(-\left(\frac{t}{\gamma}\right)^{\alpha}\right)$$

therefore,

$$h_0\left(t\right)=\frac{f\left(t\right)}{S\left(t\right)}=\frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}$$

3. Parameter Estimation Of Cox Regression

Based on Cox [12], the suggested estimation method is using MPLE, with partial likelihood function as follows:

$$PL\left(\boldsymbol{\beta}\right)=\prod_{i=1}^{n}\left(\frac{\exp\left(\boldsymbol{\beta}^T\mathbf{X}\right)}{\sum_{j=R_i}\exp\left(\boldsymbol{\beta}^T\mathbf{X}\right)}\right)^{\delta_i} \tag{2}$$

where i is a subject of T, and R_i is a risk factor of T. MPLE in the cox regression model results in a unclosed form parameter estimation, and to obtain the result of closed form estimation, newton raphson iteration method is used with equation:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{c+1} &= \boldsymbol{\beta}^c+\mathbf{s}^c\\ &= \boldsymbol{\beta}^c-\left(\mathbf{H}\left(\boldsymbol{\beta}\right)^c\right)^{-1}\mathbf{g}\left(\boldsymbol{\beta}\right)^c\end{aligned}$$

$\mathbf{g}\left(\boldsymbol{\beta}\right)$ is a gradient vector containing the first derivative of the natural logarithm likelihood function, and $\mathbf{H}\left(\boldsymbol{\beta}\right)$ is a hessian matrix containing the second derivative of natural logarithm partial likelihood. Here are the vector elements $\mathbf{g}\left(\boldsymbol{\beta}\right)$ and matrix $\mathbf{H}\left(\boldsymbol{\beta}\right)$:

$$\mathbf{g}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_{i=1}^{n}\left(x_{i1} - \frac{\sum_{l=R_i} x_{il1}\exp\left(\sum_{j=1}^{p}\beta_l x_{ij}\right)}{\sum_{l=R_i}\exp\left(\sum_{j=1}^{p}\beta_l x_{ij}\right)}\right) \\ \sum_{i=1}^{n}\left(x_{i2} - \frac{\sum_{l=R_i} x_{il2}\exp\left(\sum_{j=1}^{p}\beta_l x_{ij}\right)}{\sum_{l=R_i}\exp\left(\sum_{j=1}^{p}\beta_l x_{ij}\right)}\right) \\ \vdots \\ \sum_{i=1}^{n}\left(x_{ip} - \frac{\sum_{l=R_i} x_{ilp}\exp\left(\sum_{j=1}^{p}\beta_l x_{ij}\right)}{\sum_{l=R_i}\exp\left(\sum_{j=1}^{p}\beta_l x_{ij}\right)}\right) \end{bmatrix}$$

$$\mathbf{H}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial^2 \ln PL(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 \ln PL(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_{j^*}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln PL(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_1} & \cdots & \frac{\partial^2 \ln PL(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_{j^*}} \end{bmatrix}$$

That iteration will reach convergent if $\left|\boldsymbol{\beta}^{c+1} - \boldsymbol{\beta}^{c}\right| \le \varepsilon$

4. The Algorithm of Parameter Estimation in Matlab Code

The estimation of cox regression parameter consists of two algorithms. The first algorithm is to get parameter estimation at baseline hazard function and the second is at exponent function.

4.1. Algorithm 1. Parameter Estimation of Baseline Hazard function

Input : T, ε

Initials : $\boldsymbol{\theta}^1 = \left[\alpha^1, \gamma^1\right]^T, diff^1$

Output : $\boldsymbol{\theta}^{c+1} = \left[\alpha^{c+1}, \gamma^{c+1}\right]^T$

Do while $\left|diff^{c+1}\right| > \varepsilon$

- Calculating gradient vector $\mathbf{g}(\boldsymbol{\theta})^c$ containing the first derivative of $\ln L\left(f(t)\right)$
- Calculating hessian matrix $\mathbf{H}(\boldsymbol{\theta})^c$ containing the second derivative of $\ln L\left(f(t)\right)$
- Computing newton rule $\mathbf{s}^c = \left(\mathbf{H}(\boldsymbol{\theta})^c\right)^{-1}\mathbf{g}(\boldsymbol{\theta})^c$
- Calculating $\hat{\boldsymbol{\theta}}^{c+1} = \hat{\boldsymbol{\theta}}^c + \mathbf{s}^c$
- Calculating $diff^{c+1} = \hat{\boldsymbol{\theta}}^{c+1} + \hat{\boldsymbol{\theta}}^c$

End

4.2. Algorithm 2. Parameter Estimation of the Regression Coefficient on the Exponent Function.

Input : T, $\mathbf{X}$, ε

Initials : $\boldsymbol{\beta}^1, diff^1$

Output : $\boldsymbol{\beta}^{c+1}$

Do while $\left|diff^{c+1}\right| > \varepsilon$

- Calculating gradient vector $\mathbf{g}(\boldsymbol{\beta})^c$ containing the first derivative of $\ln PL(\boldsymbol{\beta})$
- Calculating hessian matrix $\mathbf{H}(\boldsymbol{\beta})^c$ containing the second derivative of $\ln PL(\boldsymbol{\beta})$
- Computing newton rule $\mathbf{s}^c = \left(\mathbf{H}(\boldsymbol{\beta})^c\right)^{-1} \mathbf{g}(\boldsymbol{\beta})^c$
- Calculating $\hat{\boldsymbol{\beta}}^{c+1} = \hat{\boldsymbol{\beta}}^c + \mathbf{s}^c$
- Calculating $diff^{c+1} = \hat{\boldsymbol{\beta}}^{c+1} + \hat{\boldsymbol{\beta}}^c$

End

Calculating

- the log-likelihood
- z score
- the standard error
- p-value
- the AIC

5. Result and Discussion

The written matlab is used to predict the hazard rate of DHF patients based on covariates age, sex, hemoglobin, leucocytes, hematocrit and thrombocyte. The sample used in this study are 200 patients aged between 1 to 72 years with patient thrombocyte condition while suffering from DHF was under 150.000 g/dl for all patients. DHF data to be simulated in the code are divided into 10-fold with each fold consists of 20 samples. The results of fold division will be grouped into 4 scenarios of training and testing data, namely 6:4, 7:3, 8:2, and 9:1.

The result of algorithm 1 is the coefficient of parameter $\boldsymbol{\theta} = [\alpha, \gamma]^T$. The algorithm 1 is a numerical method to solve the unclosed form parameter estimation case of the weibull distribution, i.e the iteration method of newton raphson with $\varepsilon = 0.05$ and the initial value for the parameter α and γ are 5 and 1 respectively. Maximum iterations and parameter estimation results from each scenario are as follows:

Table 1. Result of each scenario

Scenario	Maximum iteration	$\boldsymbol{\theta} = [\alpha, \gamma]^T$
6:4	31	$[2.2498, 0.0462]^T$
7:3	33	$[2.2392, 0.0455]^T$
8:2	37	$[2.2343, 0.0452]^T$
9:1	41	$[2.2349, 0.0452]^T$

The second algorithm is to get parameter estimation on regression coefficient. As in algorima 1, in the algorithm 2 the written code is a numerical method for obtaining parameter estimation on the unclosed form model coefficients with $\varepsilon = 0.0001$ and the initial value for the parameter $\boldsymbol{\beta} = (0,0,0,0,0,0)^T$ and the maximum iteration of scenario 6 : 4, 7 : 3, 8 : 2, and 9 : 1 are 30, 37, 41, and 44 respectively. The results obtained from this algorithm are:

Table 2. Parameter estimation of cox regression for each scenario

Covariates	6:4		7:3		8:2		9:1	
	beta	p-value	beta	p-value	beta	p-value	beta	p-value
Age	-0.0210	0.0000*	-0.0215	0.0000*	-0.0221	0.0000*	-0.0227	0.0000*
Sex	0.0879	0.6690	0.1677	0.3659	0.1631	0.3445	0.1530	0.3498
Hemoglobin	0.0241	0.6010	0.0169	0.7009	0.0292	0.4785	0.0322	0.4061
Leukocytes	0.0000	0.6387	0.0000	0.8900	0.0000	0.7583	0.0000	0.9789
Hematocrit	-0.0017	0.9252	0.0085	0.6068	0.0118	0.4407	0.0154	0.3013
Thrombocyte	0.0000	0.0244*	0.0000	0.0124*	0.0000	0.0032*	0.0000	0.0036*

It can be seen from Table 2 that significant covariates are age and platelets in all scenarios. It is based on p-value on age and thrombocyte which is less than $\alpha = 0.05$. The results of this prediction are in line with research conducted by Ju and Brasier [13] and Mallhi et al [14].

Ju and Brasier [13] in 2013 conducted a research on the selection of variables affecting DHF by using the classification method. The methods used are Multivariate Adaptive Regression Spline (MARS), Learning Ensemble, Random Forest (RF), Bayesian Moving Averaging (BMA), Stochastic Search Variable Selection (SSVS), Generalized Regularized Logistic Regression. The results show that the variables affecting DHF are IL-10, thrombocyte, and lymphocyte with the highest accuracy in Generalized Regularized Logistic Regression method. On the other hand, Mallhi et al [14], examined the differences in the characteristics of laboratory results between DF and DHF. The results of the study using multivariate regression analysis showed that one of the some significant factors affecting DHF is age.

The results of survival estimation of the data used for each scenario can be seen in the figures below

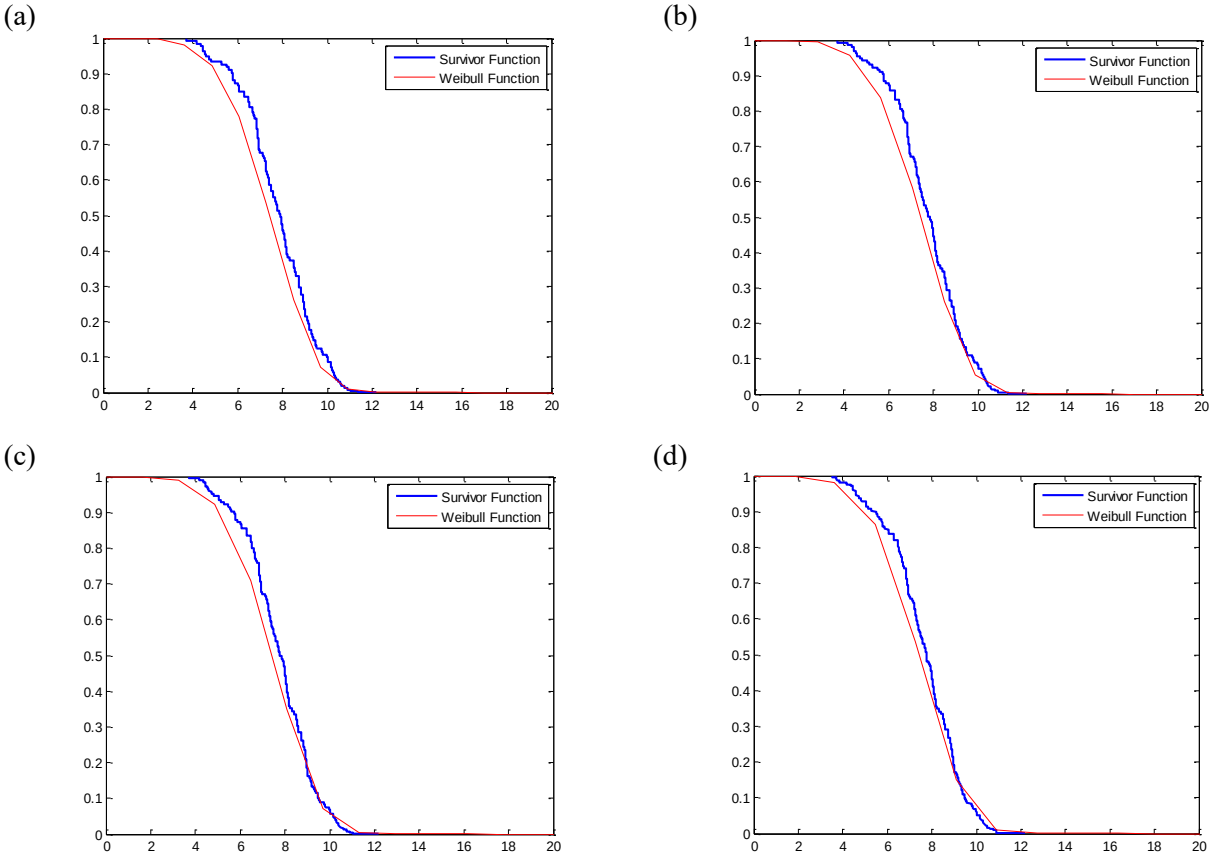

Figure 1. Result of survival estimation of four scenarios: scenario 6:4 (a); scenario 7:3 (b); scenario 8:2 (c); scenario 9:1

The figures above are an explanation of the chances of patients being able to survive when exposed to dengue. For example figure 1, patients treated for 4 days are predicted 99% to survive and recover. While the the patients were treated more than 4 days, the percentage for healing was lower, as patients treated for 10 days were only about 10%, and they who were treated more than 10 days had the lower percetage than 10%. In Figure 2, scenario 7:3 , patients treated over 10 days were estimated to be able to survive and recover from DHF less than 8%, while the patients in the scenario 8:2 are under 5.5%, and the figure for scenario 9:1 is less than 4.5%. Based on these explanations, it can be seen that patients in the scenario 6:4 have higher probability to survive and to recover from DHF compared to patients in the scenario 7:3, 8:2, and 9:1.

The results of prediction in the table 2 do not show significant covariate changes to the model. However, the predicted outcomes of each scenario can be known by using the statistical test of akaike information criterion (AIC). The smallest AIC value is the best model for predicting the survival time of DHF sufferers. As Tabatabai, et al [15], use the AIC to determine the goodness of the model. AIC in the study is used to determine the appropriate model in the predicted survival time in patients with breast cancer.

In this study, the model that is suitable for the prediction of survival time is in scenario 6: 4 which has AIC value 455.13. AIC for each scenario can be seen in table 3

Table 3. Goodness of fit for each scenario

Skenario	-2(loglikelihood)	AIC
6:4	443.13	455.13
7:3	632.97	644.97
8:2	537.93	549.93
9:1	731.66	743.66

Model formed from scenario 6:4 with the significant covariates as follows:

$$h\left(t|X\right) = h_0\left(t\right)\exp\left(\boldsymbol{\beta}^T\mathbf{X}\right)$$
$$= \frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}\exp\left(-0.021x_1 + 0.000x_6\right)$$

x_1 is age, and x_6 is thrombocyte. The model can be interpreted by using Hazard Ratio (HR). For instance, the effect of age on patients affected by DHF between patient aged 70 years and 10 years. The results is as follow:

$$HR = \frac{h\left(t, x=70\right)}{h\left(t, x=10\right)} = \frac{h_0\left(t\right)\exp\left(-0.021\times 70\right)}{h_0\left(t\right)\exp\left(-0.021\times 10\right)} = 0.284$$

HR between them is 0.284. This suggests that patients aged 70 years need a longer time to recover compared to patients aged 10. In other words, the recovery rate of patients aged 10 years is 4 times faster than patients aged 70 years.

The second interpretation is the effect of thrombocyte on the healing rate of patients with DHF. The patients thrombocyte between 10,000 g / dl and 90,000 g / dl are compared. The result is:

$$HR = \frac{h\left(t, x=10000\right)}{h\left(t, x=90000\right)} = \frac{h_0\left(t\right)\exp\left(0.000\times 10000\right)}{h_0\left(t\right)\exp\left(0.000\times 90000\right)} = 1$$

The patients thrombocyte of 10,000 g / dl and 90,000 g / dl results in a ratio of 1. This indicates that between patients who have 10,000 g/dl have the same cure rate as those who have 90,000 g/dl. This is possible to occur, because the patients thrombocyte is below 150,000 g/dl, i.e the lower limit of normal thrombocyte. 200 patients treated had levels of trhombocite are below normal, the lowest was 10.400 g/dl and 93.900 g/dl for the highest.

6. Conclusion

The simulation of the matlab code of the cox model for predicting survival time reaches convergent with the number of iterations less than 50 for each algorithm and for all scenarios. The results show that the best model is in scenario 6:4 with AIC value at 455.13 which has maximum iteration 31 in algorithm 1 and 30 in algorithm 2. Covariate affecting the healing rate of patients are age and thrombocyte.

Acknowledgments

We are grateful to Ministry of Research, Technology, and Higher Education of the Republic of Indonesia for fully funding the research in collaborative research grants between universities in Indonesia.

References

[1] Ahmed, F. E., Vos, P. W., & Holbert, D. 2007. *Modeling survival in colon cancer : a methodological review*, 12, 1–12.

[2] Omurlu, I. K., Ture, M., & Tokatli, F. 2009. Expert Systems with Applications The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. *Expert Systems With Applications*, 36(4), 8582–8588.

[3] Ihwah, A. 2015. The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product. *Italian Oral Surgery,* 3, 78–83.

[4] Omurlu, I. K., Ozdamar, K., & Ture, M. Expert Systems with Applications Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets. *Expert Systems With Applications,* 36(8), 11341–11346.

[5] Side, S., Rangkuti, Y.M., Pane, D.G., Sinaga, M.S., 2017. SEIR model simulation for Hepatitis B. *AIP Conference Proceedings*, 1885(1).020198.

[6] Side, S., Mulbar, U., Sidjara, S., Sanusi, W., 2017. A SEIR model for transmission of tuberculosis. *AIP Conference Proceedings*, 1830, 020004.

[7] Side, S., 2015. A susceptible-infected-recovered model and simulation for transmission of tuberculosis. *Advanced Science Letters.* 21(2), 137-139.

[8] Devarajan, K., & Ebrahimi, N. 2011. A semi-parametric generalization of the Cox proportional hazards regression model: inference and applications. *Comput. Statist. Data Anal*., 55, 667–676.

[9] Gradowska, P. L., & Cooke, R. M. 2011. Least squares type estimation for Cox regression model and specification error. *Computational Statistics and Data Analysis*, 56(7), 2288–2302.

[10] Ojeda, F. M., Muller, C., Bornigen, D., Tregouet, D., Schillert, A., Heinig, M., Zeller, T., & Schnabel, R. B. 2016. Comparison of Cox Model Methods in a Low-dimensional Setting with Few Events. 14 , 235–243.

[11] Chen, M., Ibrahim, J. G., & Shao, Q. Maximum likelihood inference for the Cox regression model with applications to missing covariates. *Journal of Multivariate Analysis*, 100(9), 2018–2030.

[12] Cox, D.R., & Oakes, D. 1984. *Analysis of Survival Data. London: Chapman & Hall*.

[13] Ju, H., & Brasier, A. R. 2013. Variable selection methods for developing a biomarker panel for prediction of dengue hemorrhagic fever. *BMC Research Notes*, 6(365), 1–8.

[14] Mallhi, T. H., Khan, A. H., Adnan, A. S., Sarriff, A., & Khan, Y. H. Clinico-laboratory spectrum of dengue viral infection and risk factors associated with dengue hemorrhagic fever : a retrospective study. *BMC Infectious Diseases*, 15(399), 1–12.

[15] Tabatabai, M. A., Eby, W. M., Nimeh, N., Li, H., & Singh, K. P. 2012. Clinical and multiple gene expression variables in survival analysis of breast cancer : Analysis with the hypertabastic survival model. *BMC Medical Genomics,* 5(63), 1–13.