Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 10, Number 4 (2014), pp. 581-590

http://www.ripublication.com

Comparative Method in PCA and PLS COX Regression to Solve Multicollinearity

Adji Achmad Rinaldo Fernandes and Solimun

Departement of Mathematics, Statistics Study Program, Universitas Brawijaya
Jl. Veteran 169 Malang 65111 Email: fernandes@staff.ub.ac.id

ABSTRACT

Cox regression is one of the method in survival analysis which is used to explain the correlation and influence between the individual failure at a certain time with one or more predictor variables. There is an assumption in Cox regression that is non multicollinierity. If non multicollinierity assumption not fulfill it means that using Cox regression is no longer appropriate. Therefore some alternatif method developed to solve this multicollinerity problem such Partial Least Square-Cox regression (PLS-Cox) [3], and Principal Component Analysis-Cox regression (PCA-Cox) [4]. The objective of this research is to observe different influence of predictor variables to individual failure probability and compare which one of the best method between PLS-Cox and PCA-Cox in survival data analysis that contains multicollinierity. This research using survival data of the probability of myocardial recovery in patients with CVA hospitalize at Rumah Sakit Haji Surabaya [14]. Subjects studied as many as 69 patients. Response variable (Y) is the length of patient hospitalization (days), with predictor variables were age (X1), gender (X2, 0 = female, 1 = male), diagnose (X3, 4 categories, complications of diabetes, complications of hypertension, more than one disease complications, and without complications), LDL cholesterol (X4), HDL cholesterol (X5), and Blood Pressure (X6).Based on the adjusted value of cross validation ($Q^2_{adjusted}$), it can be concluded that modeling using PLS-Cox regression on data containing multicollinearity can provide a good prediction capability and relatively better.

Keywords: Survival, PLS-Cox, PCA-Cox, Multicollinearity

Mathematics Subject Classification: 62N03

1. INTRODUCTION

Survival analysisis onestatisticalanalysis techniquethat discuss the datarelating to

thefailureeventat a time fromthe initialconditions ofthe studyuntil theeventends. Survivalanalysisortest analysisis the studyof livinga longtime in betweenevents. The results obtainedare Twhichis definedas the timean eventoccurs, commonly referred to assurvival timeorsurvivaltime[1].

One method that can be used for modeling the response variables in the form of survival time with one or more predictor variables are the Cox Proportional Hazard Regression model or commonly referred to by Cox regression model [2]. If the Cox regression analysis involving more than one predictor variable, it will be very possible there is a close relationship between the predictor variables or characterized by multicollinearity, so the use of Cox regression is no longer appropriate. According to Bastien [3] one of the methods can be used to overcome the multicollinearity problem in Cox regression is to use the PLS-Cox method (Partial Least Square-Cox Regression). PLS-Cox is a method that can be used to establish and improve the ability of predictive models when non-multicollinearity assumption is not met. In addition to PLS method proposed by Bair, et al.[4] states that the Principal Component Analysis Cox Regression (PCA-Cox) is one of the alternative methods that can be used to overcome the problem of multicollinearity.

Based on the above, in this study will apply two methods to overcome multicollinearity on the Cox regression method of Cox Principal Component Regression (PCA-Cox) and Partial Least Square method-Cox Regression (PLS-Cox), otherwise it will be determined which method best among them to overcome the multicollinearity in the Cox regression survival analysis.

2. LITERATURE REVIEW

2.1. Survival Function and Cox Regression Model

Suppose t is defined as the actual survival time of an object, and the value of the variable T has a non-negative value. Survival function S (t), defined as an opportunity or an object has probabiltas survival time greater than t, in other words an object has a chance of living longer than t can be expressed as [5]:

$$S(t) = P(T \geq t) = 1 - F(t) \quad (1)$$

The basis of Cox regression model generated from the hazard function for the i-th object at the time t that consists of two factors: the baseline hazard function and symbolized as a linear function of a set of predictor variables that were generated by the exponent. In general Cox regression model is defined as follows [5]

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + ... + \beta_k x_{ik}) \quad (2)$$

$h_i(t)$ is an opportunity to object-i failure or death at time t, is the baseline hazard function, and $h_0(t)$ is the density function of resistance opportunity to object-t, is the survival function. Hazard function values, determined after the value is obtained.

Estimation of parameters in the Cox regression model with maximum likelihood approach. According to Bastien (2004) likelihood functions for Cox Regression model are:

$$L(\beta^{'}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right\}^{\delta_i}, \quad (3)$$

where $R(t_i)$ is a group of objects that are at risk t_i, and δ_i is a censored indicator which is zero if t_i, i = 1,2, .., n is the right censored and valuable one for the other. Parameter estimation of β in Cox Regression models obtained by maximizing the log-likelihood function using the Newton-Raphson procedure in which the estimation of parameters β_1, β_2,..., β_p obtained from the completion of a x_{1p} vector equations are expressed by a score of coefficient vectors.

2.2. Multicollinearity Assumption in Cox Regression Model

Bastien et al. [6] stated that the Cox regression modeling involving more than one predictor variable must meet the multicollinearity assumption that information obtained from the modeling results are not biased.The use of the term applied by Ragnar Frisch Multicollinearity. Multicollinearity is a linear relationship between several predictor variables that occur in the model. multicollinearity is used to show a linear relationship between several predictor variables in the regressionmodel.

According to Nguyen and Rocke [7], Cox regression is very sensitive to the presence of multicollinearity or linear relationship between predictor variables. This is because the data is the Cox regression survival data were used to analyze several predictor variables on the response variable (survival time) by the number of observations (number of objects) is limited.One way that can be used to detect the presence of multicollinearity between predictor variables is to look at the value of VIF. The existence of multicollinearity if the VIF values> 5 indicate the presence of multicollinearity [8]

There are some impacts due to multicollinearity. The first, variant Cox regression coefficient becomes large. The second figure would have obtained estimates of the value that is inconsistent with the substance, so it can be misleading interpretation. According to Gujarati [9], multicollinearity causes the estimates of regression coefficients become unstable. This is due to the large standard deviation, so that the accuracy of estimates of regression coefficients would be difficult to obtain.

One method that can be used for modeling the response variables in the form of survival time with one or more predictor variables are the Cox Regression model or commonly referred to by Cox regression model [2]. If the Cox regression analysis involving more than one predictor variable, it will be very possible there is a close relationship between the predictor variables or characterized by multicollinearity, so the use of Cox regression is no longer appropriate. According to Bastien [3], one of the methods can be used to overcome the multicollinearity problem in Cox regression is to use the PLS-Cox method (Partial Least Square-Cox Regression). PLS-Cox is a method that can be used to establish and improve the ability of predictive models when non-multicollinearity assumption is not met. In addition to PLS method proposed by Bastien [3], Bair, et al. [4] states that the Principal Component Analysis Cox Regression (PCA-Cox) is one of the alternative methods that can be used to overcome the problem of multicollinearity.

2.3. Partial Least Square Cox Regression (PLS-Cox) Model

Partial Least Square Regression (PLS) is an established method of linear regression equation by forming a new predictor variable is usually called by a factor or component that functions as a linear combination of original variable says X [10]. According to Jian et al. [11], the PLS method to reduce the variables into a variable or a new component with many components with a number less than or equal to the smallest dimension matrix of predictor variables. The new component has formed an independent component or not, there is a correlation between each other, so it can be used to overcome the non-fulfillment of the assumption of non-multicollinearity on regression.

According Boulesteix and Strimmert [12], PLS regression decomposition is based on the response variable and predictor variables. Matrix of predictor variables X = $\{x_1, x_2, ..., x_p\}$ is a matrix that determines the survival time and serves to form an orthogonal PLS components K_h. In PLS-Cox regression K_h formation component for h = 1, 2, ..., m can be written as an equation:

$$K_h = w_{1h}x_1 + w_{2h}x_2 + \ldots + w_{ph}x_p \quad (4)$$

Where $w_{jh}$is the coefficient normalization of a_{jh}

$$\mathbf{w}_{\mathbf{jh}} = \frac{\mathbf{a}_{\mathbf{jh}}}{\left\|\mathbf{a}_{\mathbf{jh}}\right\|} \quad (5)$$

The coefficient of a_{jh} is a Cox regression coefficients for predictor variables. The normalized coefficient to the PLS using only the significant coefficient of a_{jh}, while for a_{jh}insignificant coefficient set equal to zero. a_{jh}in the Cox regression coefficient is written as:

$$h_i(t) = h_0(t) \exp(c_1k_1 + c_2k_2 + \ldots + c_{h-1}k_{h-1} + a_{jh}x_{jh}) \quad (6)$$

where $x_{jh}$is avector ofresidualsobtainedfromregressionbetween$x_{j(h-1)}$withk_1, ..., k_{h-1}.

Establishment procedures component k_h for h = 1, 2, ..., m on Cox regression can be described as follows:

1. Calculate the first componentk_1 (h = 1)

a) Obtain the coefficient $a_{j1}$from the model of regression x_{j1}or x_j (initial predictor variable) on Ypartially, so we get the equation:

$h_i(t) = h_0(t) \exp(a_{11}x_{11})$

$h_i(t) = h_0(t) \exp(a_{21}x_{21})$

$\vdots \qquad\qquad \vdots$

$h_i(t) = h_0(t) \exp(a_{p1}x_{p1})$

b) Calculate the normalized coefficient of w_{j1} as follow:

$w_{j1} = 0$ if $a_{j1}$is not significant

$$\mathbf{w}_{\mathbf{j1}} = \frac{\mathbf{a}_{\mathbf{j1}}}{\left\|\mathbf{a}_{\mathbf{j1}}\right\|} \quad \text{if } a_{j1}\text{is significant}$$

c) Calculate the component score $k_1$from the equation:

$k_1 = w_{11}x_1 + w_{21}x_2 + \ldots + w_{p1}x_{p1}$

2. Calculate the h-th component k_h(h = 2,3, …, m)

a) Calculate the residual vector $x_{jh}$from the regression of $x_{j1}$with k_1, k_2... k_{h-1}.

$x_{1(h-1)} = p_{11}k_1 + p_{12}k_2 + ... + p_{1(h-1)}k_{(h-1)} + x_{1h}$

$x_{2(h-1)} = p_{21}k_1 + p_{22}k_2 + ... + p_{2(h-1)}k_{(h-1)} + x_{2h}$

$\vdots \qquad\qquad \vdots$

$x_{p(h-1)} = p_{p1}k_1 + p_{p2}k_1 + ... + p_{p(h-1)}k_{(h-1)} + x_{ph}$

b) Obtain the coefficient $a_{jh}$from the model of regression $x_{jh}$and k_1, k_2, ... k_{h-1}on Ypartially, so we get the equation:

$h_i(t) = h_0(t) \exp (c_1k_1 + c_2k_2 + ... + c_{(h-1)}k_{(h-1)} + a_{1h}x_{1h})$

$h_i(t) = h_0(t) \exp (c_1k_1 + c_2k_2 + ... + c_{(h-1)}k_{(h-1)} + a_{2h}x_{2h})$

$\vdots \qquad\qquad \vdots$

$h_i(t) = h_0(t) \exp (c_1k_1 + c_2k_2 + ... + c_{(h-1)}k_{(h-1)} + a_{ph}x_{ph})$

c) Calculate the normalized coefficient of w_{jh}as follow:

$w_{jh} = 0$ if $a_{jh}$is not significant

$$\mathbf{w_{jh}} = \frac{\mathbf{a_{jh}}}{\left\|\mathbf{a_{jh}}\right\|} \qquad \text{if } a_{jh} \text{is significant}$$

d) Calculate the component score $k_h$from the equation:

$k_h = w_{1h}x_{1h} + w_{22}x_{2h} + \ldots + w_{ph}x_{ph}$

The k_h component regression on PLS method, by performing regression analysis between the$k_i$with Y. In other words using k_i components as predictor variables for Cox regressed with the response variable Y. The results of the regression equation k_i component of the response variable Y can be written as:

$$h_i(t) = h_0(t) \exp (c_1k_1 + c_2k_2 + \ldots + c_hk_h) \qquad (7)$$

The equation above is a Cox regression model with the predictor variable component obtained using the PLS method, so briefly above equation can be called a model of the Partial Least Square Cox Regression or PLS-Cox.

2.4. Principal Component Analysis Cox Regression (PCA-Cox) Model

According to Bair et al [4], PCA can be applied to the problem of multicollinearity in regression analysis such as Cox regression in survival analysis. PCA method can also see the influence of predictor variables and to identify which variables are important predictors. In principal component analysis, the column vector of p variables X origin is transformed into a column vector q as a new variable K where $q \leq p$. In equation form, PCA is expressed as:

$$k_i = \sum_{j=1}^{p} a_{ij}X_j \qquad (8)$$

where $1 \leq i \leq q$, $1 \leq j \leq p$ and $k_1, k_2, \ldots, k_q$ independently, where q is the number of component as new variables depend on the proportion of cumulative variance greather than 75%, or eigen value more than 1.The k_i component regression on PCA method, by performing regression analysis between the$k_i$with Y. In other words using k_i components as predictor variables for Cox regressed with the response variable Y. The results of the regression equation k_i component of the response variable Y can be written as:

$$h_i(t) = h_0(t) \exp (c_1k_1 + c_2k_2 + \ldots + c_qk_q) \qquad (9)$$

The equation above is a Cox regression model with the predictor variable component obtained using the PCA method, so briefly above equation can be called a model of the Principal Component Analysis Cox Regression or PCA-Cox.

2.5. Model Predictivity

Model predictive with the value of cross validation (Q^2) can be used to determine how better the accuracy of predictions generated from the model formed [13]. Comparison of PLS-Cox regression model and PCA-Cox regression model can be done by comparing the adjusted Q^2 values obtained from both models. This is because the value of an adjusted Q^2 is a measure of predictive ability that considers the number of predictor variables which are formed in the model. The formula of Q^2 adjusted can be obtained by the following equation:

$$Q^2\text{adjusted} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 \,/\, n-k}{\sum_i (y_i - \bar{y})^2 \,/\, n} \tag{37}$$

Where k is the number of component, y_i is the i-th response variable, $\hat{\mathbf{y}}_i$ is the predictive value of y_i, and $\bar{\mathbf{y}}$ is the mean of y_i. The value of $Q^2_{adjusted}$ between -1 to 1, where the value closes to 1 indicate the predictions are more accurate.

3. ANALYSIS RESULT

3.1. Data Representation

This research using survival data of the probability of myocardial recovery in patients with CVA hospitalize at Rumah Sakit Haji Surabaya [14]. Subjects studied as many as 69 patients. Response variable (Y) is the length of patient hospitalization (days), with predictor variables were age (X1), gender (X2, 0 = female, 1 = male), diagnose (X3, 4 categories, complications of diabetes, complications of hypertension, more than one disease complications, and without complications), LDL cholesterol (X4), HDL cholesterol (X5), and Blood Pressure (X6).

The results of testing the assumption of multicollinearity in a row gives the value of VIF 1.0, 1.1, 1.1, 13.4, 9.8, and 4.2 indicating that multicollinearity assumptions are not fulfiled. It means that using Cox regression is no longer appropriate, and PLS-Cox and PCA are the alternative methods that can be used to overcome the problem of multicollinearity.

3.2. The Result of PLS-Cox Model

In PLS-Cox, the calculation of the formation of a compact component is shown in Table 1 below:

Tabel 1. Calculation of component PLS-Cox

Component	Equation
k_1	$k_1 = 0.69667\ x_4 + 0.60122\ x_5 + 0.39139\ x_6$
k_2	$k_2 = 0.87354\ x_4 + 0.39626\ x_5 - 0.28295\ x_6$
k_3	$k_3 = 1.0000\ x_1$
k_4	$k_4 = -3.61612\ x_{3.2} + 0.99007\ x_4 - 0.59743\ x_5 + 1.29290\ x_6$
k_5	not formed

Table 1 shows that only four components are formed on the PLS-Cox k_1, k_2, $k_3$and k_4, while componentk_5 is not formed because there was no significant a_{j5} coefficients. Then the formation of PLS-Cox regression model to proceed with regress the four components with a variable Y (length of stay of patients) so that the results of Cox regression coefficients obtained are shown in Table 2 as follows:

Table2. The Result of PLS-Cox

Component	Coefficient	SE	Wald	p-value	95.0% Normal CI	
					Lower	Upper
k_1	0.02504	0.001	15.77	0.000	0.022	0.028
k_2	0.03913	0.007	5.34	0.000	0.024	0.053
k_3	0.00880	0.003	2.58	0.010	0.003	0.015
k_4	0.08160	0.030	2.69	0.007	0.022	0.141

Based on the value of the coefficient in Table 2, the components can be used as the basis for the Cox regression model building by incorporating the results of the regression coefficients of the components k_1, k_2, k_3 and k_4 with a variable Y, so we get the equation PLS-Cox regression model in the form of original variable X as the following:

$h_i(t) = h_0(t) \exp (0{,}00880\ x_1 - 0{,}29508\ x_{3.2} + 0{,}13241\ x_4 - 0{,}01819\ x_5 + 0{,}10423\ x_6)$

Significance test parameters PLS-Cox regression model using bootstrapping method, the result shown in Table 3 as follows

Table 3: Results of Significance Testing Parameter PLS-Cox Regression Model

Variable	Coefficient	p-value
X_1	0.00880	0.94571
X_2	not formed	
$X_{3.1}$	not formed	
$X_{3.2}$	-0.29508	0.99270
$X_{3.3}$	not formed	
X_4	0.13241	0.00000
X_5	-0.01819	0.00000
X_6	0.10423	0.00000

From the result shows that of only five predictor variables are age (X_1), Diagnose of Complications Hypertension ($X_{3.2}$), LDL cholesterol (X_4), HDL cholesterol (X_5) and Blood Pressure (X_6) that form the regression model PLS-Cox. From the five variables that form, only LDL cholesterol variable (X_4), HDL cholesterol (X_5) and Blood Pressure (X_6) that provide a significant influence on length of stay of patients with CVA infarction hospitalize at Rumah Sakit Haji Surabaya.

3.3. The Result of PCA-Cox Model

In PCA-Cox, the calculation of the formation of a compact component is shown in Table 4 below:

Tabel 4. Calculation of component PCA-Cox

Component	Equation
k_1	$k_1 = 0{,}059\ x_1 + 0{,}172\ x_2 - 0{,}157\ x_{3.1} - 0{,}013\ x_{3.2} - 0{,}120\ x_{3.3} - 0{,}569\ x_4 - 0{,}555\ x_5 - 0{,}544\ x_6$
k_2	$k_2 = 0{,}600\ x_1 + 0{,}237\ x_2 - 0{,}511\ x_{3.1} + 0{,}445\ x_{3.2} + 0{,}331\ x_{3.3} + 0{,}055\ x_4 + 0{,}094\ x_5 + 0{,}052\ x_6$
k_3	$k_3 = - 0{,}004\ x_1 + 0{,}134\ x_2 + 0{,}228\ x_{3.1} + 0{,}681\ x_{3.2} - 0{,}678\ x_{3.3} + 0{,}056\ x_4 + 0{,}063\ x_5 - 0{,}013\ x_6$
k_4	$k_4 = - 0{,}244\ x_1 + 0{,}831\ x_2 + 0{,}358\ x_{3.1} + 0{,}037\ x_{3.2} + 0{,}334\ x_{3.3} + 0{,}036\ x_4 + 0{,}072\ x_5 - 0{,}053\ x_6$

Table 4 shows that PCA-Cox regression modeling to form three components, namely k_1, k_2, $k_3$and k_4. Each component that is formed is a linear combination of variables so that the origin of each component that is formed consisting of all of the original predictor variables. Subsequently formed the PCA-Cox regression model with regress the component with a variable Y (length of stay of patients) as follows:

Table5. The Result of PCA-Cox

Component	Coefficient	SE	Wald	p-value	95.0% Normal CI	
					Lower	Upper
k_1	-0.29255	0.022	-13.25	0.000	-0.336	-0.249
k_2	0.06224	0.042	1.49	0.137	-0.019	0.144
k_3	0.08348	0.059	1.4	0.161	-0.033	0.200
k_4	0.01168	0.043	0.27	0.787	-0.073	0.096

Based on the value of the coefficient in Table 5, the components can be used as the basis for the Cox regression model building by incorporating the results of the regression coefficients of the components k_1, k_2, k_3 and K_4 with a variable Y, so we get the equation PCA-Cox regression model in the form of original variable X as the following:

$h_i(t) = h_0(t) \exp (0{,}01690\ x_1 - 0{,}01467\ x_2 + 0{,}03735\ x_{3.1} + 0{,}08878\ x_{3.2} + 0{,}00301\ x_{3.3} + 0{,}17498\ x_4 - 0{,}17432 x_5 + 0{,}16068\ x_6)$

Significance test parameters PCA-Cox regression model using bootstrapping method, the result shown in Table 6 as follows

Table 6: Results of Significance Testing Parameter PCA-Cox Regression Model

Variable	Coefficient	p-value
X_1	0.01690	0.67137
X_2	-0.01467	0.26056
$X_{3.1}$	0.03735	0.34340
$X_{3.2}$	0.08878	0.58132
$X_{3.3}$	0.00301	0.31960
X_4	0.17498	0.00000
X_5	-0.17432	0.00000
X_6	0.16068	0.00000

From the result shows that of only three predictor variables are LDL cholesterol variable (X_4), HDL cholesterol (X_5) and Blood Pressure (X_6) that provide a significant influence on length of stay of patients with CVA infarction hospitalize at Rumah Sakit Haji Surabaya.

3.3. Comparison of PLS-Cox and PCA-Cox

The results of comparison of the influence of predictor variables on the probability of failure between the PLS-Cox regression modeling and PCA-Cox regression each of the research data are shown as Berkut:

Tabel 7. Comparison of PLS-Cox and PCA-Cox

Variable	PLS-Cox		PCA-Cox	
	Coefficient	*p-value*	Coefficient	*p-value*
X_1	0.00880	0.94571	0.01690	0.67137
X_2			-0.01467	0.26056
$X_{3.1}$			0.03735	0.34340
$X_{3.2}$	-0.29508	0.99270	0.08878	0.58132
$X_{3.3}$			0.00301	0.31960
X_4	0.13241*	0.00000	0.17498*	0.00000
X_5	-0.01819*	0.00000	-0.17432*	0.00000
X_6	0.10423*	0.00000	0.16068*	0.00000
Q^2_{adj}	0,8991		0,8379	

* significant with 5% level of significance

Based on Table 7 above, by using PLS-Cox regression model and Cox regression on PCA-significance level (α) by 5%, there are similarities predictor variable influence on survival chances. Based on Table 8 above shows that the variables have the same effect on the survival chances of CVA infarct patients is variable LDL cholesterol (X4), HDL cholesterol (X5) and Blood Pressure (X6). The similarity in this case is also seen from the side of coefficient significance and magnitude of the coefficients (a sign of the coefficient).To compare the results of which one is better between the PLS and PCA-Cox-Cox used the value of $Q^2_{adjusted}$. In the above table shows that modeling using PLS-Cox regression model showed better results than using PCA-Cox regression model. This indicates that modeling using PLS-Cox regression on data containing multicollinearity can provide a good prediction capability and relatively better.

4. DISCUSSION AND CONCLUSION

Based on the results of research using the adjusted value of cross validation ($Q^2_{adjusted}$), it can be concluded that modeling using PLS-Cox regression on data containing multicollinearity can provide a better prediction capability and relatively better. Based on this research, the recommendation for further research is suggested to be PLS-Cox regression modeling and PCA-Cox on survival data that does not meet the proportional hazards assumption by applying the PLS-Cox regression modeling and PCA-Cox on a stratified Cox model or the Cox regression model with a variable time-dependent.

5. ACKNOWLEDGEMENT
We would like to express our sincere gratitude to University of Brawijaya for providing fund for this research as part of the PhD Research Program.

6. REFERENCES

[1] Kleinbaum, D.G. and Klein, M. 2005. *Survival Analysis : A Self-Learning Text*. Second Edition. Springer-Verlag. New York.
[2] Fox, J. 2002. *Cox Proportional Hazards Regression for Survival Data*. http://www.google.com/Search:Cox-ph-Fox/files,
[3] Bastien, P. 2004. *PLS-Cox model: Application to Gene Expression*. In: COMSTAT, Section: Partial Least Square.
[4] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. 2004. *Prediction by Supervised Principal Component*. Department of Statistics and Health, Research Policy. University of Stanford.
[5] Collet, D. 2003. *Modelling Survival Data in Medical Research Second Edition.*Chapman and Hall. London.
[6] Bastien, P.,Vinci. E., dan Tenenhaus, M. 2005. *PLS Generalized Linear Regression*. Computational Statistics & Data Analysis.
[7] Nguyen, D.V. dan D. Rocke. 2002. *Partial Least Square Proportional Hazard Regression for Applicaation to DNA Microarray Survival Data*. Bioinformatics, 18, 1625-1632.
[8] Kutner, M. H., Machtseim, dan J. Neter. 2004. *Applied Linier Regression Models*. Fourth Edition. Mc Graw Hill. New York.
[9] Gujarati, D. 1995. Basic Econometric. Terjemahan: Zain, S. Erlangga, Jakarta.
[10] Li, H. dan J. Gui. 2004. *Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data*. Center for Bioinformatics & Molecular Biostatistics. University of California. San Fransisco.
[11] Jian, J.D; L. Linh, dan D. Rocke. 2006. *Dimension Redustion for Classification With Gene Expression Microarray Data*. http://www.cipic.ucdavis.edu~dmrockepapersDai_Lieu_ Rocke_SAGMB2006.pdf.
[12] Boulesteix, A. dan K. Strimmer. 2005. *Partial Least Square : A Versatile Tool For The Analysis of High-Dimensional Genomic Data*. http://www.stat.unimuenchen.desfb386. papersdspaper457.pdf.
[13] Polanski, J; A. Bak; R. Gieleciak dan T. Magdzdiarz. 2004. *Self-Organizing Neural Network for Modelling Robust 3D and 4D QSAR: Application to Dihydofolate Reductase Inhibitors*.http://www.mdpi.org/molecules/papers/91201148.pdf.
[14] Ismiyati. 2007. Application ofKaplanMeierforCVAPatientsKnowing theprobability ofhealinginfarction(Case Studyin RS Haji Surabaya). Universitas Airlangga. Surabaya. Thesis.