Applied Ridge and LASSO Methods in Cox Proportional Hazard Modelling

Garry Rusmadi, Asep Saefuddin, Bagus Sartono

Abstract— Cox Proportional Hazard (PH) is used to estimate the relationship between some variables of failure in an examination. We assess the survival for the patients of cardiovascular disease in the ICCU RSMY Bengkulu. There are 10 variables used in order to predict survival in patients, taken from the respondent demographic and laboratory test results. The more explanatory variables are used, causing correlations among variables (multicollinearity). Multicollinnearity Become a problem in data analysis Because It causes less precise results. The use of ridge regression and LASSO are Able to overcome the problem of multicollinearity. In this study, using Cox PH when there are three variables that significantly affect the survival of Patients with cardiovascular disease. The results of Cox LASSO, also indicates three variables affect the survival significantly, while the other variables become selected (0). The results of Cox Ridge only Able to shrink variable close to 0. The smallest error test using RMSEP, Indicating that the method of Cox Ridge is the best estimation models in predicting survival in patients with cardiovascular disease.

Index Terms— cox proportional hazard, cox ridge, LASSO, survival analysis.

1 INTRODUCTION

The incidence of cardiovascular disease that increase from year to year makes these disease become the number one cause of death in the world. According to the World Health Organization (WHO) (2005), 17.5 million (30%) of the 58 million deaths per year in the world caused by this disease. Based on the data that has been collected by WHO, the estimation deaths from cardiovascular disease increased to 23.6 million people in 2030. People can predict the number of incidence in the future, but the approximate length of survival time that patients owned still difficult to predict.

One of the statistical techniques used to analyze the survival time is survival analysis. In survival analysis modeling, time is the response variable. As for the other factors that affect the time of the explanatory variables. A period until an occurrence in the analysis of endurance known survival time.

Survival analysis is a statistical method used to describe the data in the form of time, ranging from observation time until the occurrence of an event. Data survival in the field of health is obtained from an observation of a group or several groups of individuals, in this case is the patient, who observed and recorded the time of the failure of any individual (Collet 2003). The failure is death due to certain diseases, recurring illness after treatment, or the emergence of new diseases. If the failure observed was the occurrence of death, then the data survival can divided into two groups:

a) Uncensored data, represents the difference between starting time doing observation until the occurrence of death.

b) Censored data, if the time of death is unknown. Represents the difference between the starting time doing observation until the last time the study.

The more explanatory variables in a model, the better the prediction model is. However, the more explanatory variables tend to lead to multikolinearity. Multicollinearity in the data causing researchers can not see the influence of independent variables on the response variable separately (Gujarati 1992). Solution for multicollinearity is to select independent variables, namely choosing independent variables which provide information on the accuracy of the prediction. Ridge regression is one of method that can be used to overcome the problem of multicollinearity. However, ridge regression can not perform the selection of independent variables to obtain the best model. Therefore, Tibshirani (1997) introduced Least Absolute Shrinkage and Operator Selection (LASSO) method. LASSO using regression techniques which do estimation by minimizing the residual sum of squares with a constraint L1 which will shrink towards zero variables coefficient and some produce precise coefficient equal to zero (Tibshirani 1996).

One of methodology for the survival analysis is Cox Prportional Hazard method. In this study, the variable selection techniques that can be performed in the process of utilizing Cox Proportional Hazard when the data consist of multicolinearity are ridge and LASSO method. In these paper, the author wants to apply ridge and LASSO methods for Cox Proportional Hazard modelling and choose the best methodology to overcome multicolinearity in data.

2 RESEARCH METHOD

2.1 Data

The data used in this research are medical records of patients with cardiovascular disease in Intencive Coronary Care Unit (ICCU) M. Yunus Hospital, Bengkulu. Patients were sampled in the study were cardiovascular diseases patients who were nursed in ICCU during January to December 2011. There are 12 variables observed in the form of 434 patients, which describe in Table 1. Data collected from patients being nursed in the ICCU until the patient leaves the ICCU room.

2.2 Methods of Data Analysis

The stages of data analysis in this research are as follow:

- *Garry Rusmadi is currently pursuing masters degree program in applied statistics in Bogor Agricultural University, Indonesia, PH +6285273666692. E-mail: gery8sc@gmail.com*
- *Asep Saefuddin is Lecturer, Department of Statistics, Bogor Agricultural University, Bogor, Indonesia*
- *Bagus Sartono is Lecturer, Department of Statistics, Bogor Agricultural University, Bogor, Indonesia*

TABLE 1
LIST FOR OF VARIABLES

Variables	Explanation
X_1	Age
X_2	Gender
X_3	Status of service
X_4	Blood sugar
X_5	Urea
X_6	Creatinine
X_7	Uric acid
X_8	Triglycerides
X_9	Low-density lipoproteins (LDL)
X_{10}	High-density lipoproteins (HDL)
Y_t	Days of nursing in ICCU
Y	Status of death (Status of cencoring)

1. Descriptive Analysis
   Descriptive analysis was performed to explore the general description of data patten that aimed to get the appropriate next analysis.
2. Checking the multicolinearity assumption
   Examine each variable by using correlation test statistic.
3. Perform Cox Proportional Hazard (PH) modelling
4. Testing Cox PH parameters by using Wald test
5. Perform Cox Ridge
6. Perform LASSO Cox
   Stages in EM algorithm, i.e:
   a. Standardise the independent variable s
   b. Determine the results of the estimation parameters.
   c. Looking forvalue of λ, the minimum with λ is aparameter tuning whose value is selected from generalized cross validation that minimum
   d. Perform the iterative process so as to obtain the value of β^convergence.
   e. To test the significance of parameter β^by using the Wald test
7. Selection of the best model and conclusion.
   Selection of the best estimation method to determine the value of the smallest error estimation method RMSEP

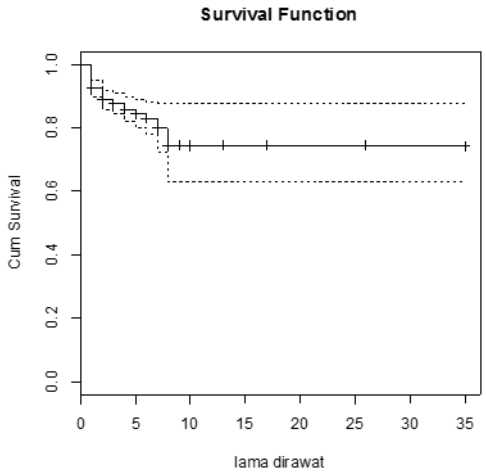
3 RESULT AND DISCUSSION

3.1 Descriptive Analysis

Survival probability of patients survive on the firts day was 0.925, and on the second day becomes 0.89 (Figure 1). The largest decline of probablity to survive occurred around the first 8 days. This means that many patients failure to alive in ICCU (death) occurred in the first 8 days.

Survival function for female patients was lower than male patients. Judging from the characteristics of patient care, patient JAMKESNAS has the lowest endurance function, while patients by insured company (PT) has the function of the highest survival.

FIGURE 1
SURVIVAL PLOT FOR CARDIVASCULAR DEASEAS IN ICCU



Furthermore, at the variable blood sugar, patients with high blood sugar have the greatest risk of death. As well as the levels of uric acid, creatinine, triglycerides, and LDL. In contrast, patients who had lower levels of urea have the greatest risk of death. This means that the higher the levels of urea, the smaller the risk of the patient dying. This also applies to the levels of HDL in the blood.

3.2 Checking Multicolinearity Assumption

Checking multikolinierity assumptions approached by checking the correlation between the variables of observation using Pearson test. Variable age has association with type of service and HDL status, variable blood sugar status linked to creatinine status, variable creatinine status associated with uric acid status, variable uric acid status associated with triglycerides status and LDL, and variables triglycerides status linked to HDL status.

3.3 Cox Proportional Hazard Modelling

The result of Cox proportional hazard models to the data survival of cardiovascular patients in ICCU, shows two variables have significant effect on alpha 10%. The variables are X3D1 (Status Services Askes) and X4 (blood sugar). More detailed information can be found in Table 2.

3.4 Application Methods on Cox Ridge

Estimation of ridge coefficient obtained from the selection of the optimal λ value. The results of ridge trace subjective in the selection of λ value. That is because the difficulty of determining the minimum value of λ when the value of β (λ) began to stabilize in each independent variable. An optimum λ can be recovered from the minimum GCV value, ie when λ of 0.4099. The coefficient estimator ridge can be seen in Table 3.

3.5 Application Methods on Cox LASSO

Coefficients estimation of LASSO obtained by computational algorithms in R using LARS. Initiation in the algorithm begins by setting all coefficients with zeros.

TABLE 2
COEFFICIENT COX PROPORTIONAL HAZARD

Peubah	Koefisien	exp(koefisien)	z	p
x1	-0.00018	0.999823	-0.02	0.984
x2D1	0.36932	1.44675	1.35	0.178
x3D1	-0.62639	0.53452	-1.82	**0.068**
x3D2	-0.07643	0.926414	-0.22	0.825
x3D3	-0.61577	0.540228	-0.81	0.418
x4	0.003127	1.003131	1.88	**0.060**
x5	-0.01398	0.986122	-0.83	0.406
x6	-0.08644	0.91719	-0.61	0.545
x7	-0.01218	0.987897	-0.63	0.527
x8	0.000387	1.000388	0.2	0.839
x9	-0.00418	0.995833	-0.25	0.802
x10	0.000855	1.000855	0.27	0.790

Furthermore, variables that have a high correlation entered one by one into the model. The first iteration X3D1 variables that have the highest correlation with other variables remnant compared to $\sum|\hat{\beta}_j| / \max \sum|\hat{\beta}_j|$ about 0.018 incoming X2D1. Variables X4 subsequent entry into the model with $\sum|\hat{\beta}_j| / \max \sum|\hat{\beta}_j|$ until 0.035. The stages of the inclusion of variables is X3D1 X2D1 X4.

3.6 Model Comparison between Cox PH, Cox Ridge, and Cox LASSO

Based on Table 3 can be seen the comparison results of modeling coefficient from the three methods. Ridge coefficient estimated value tends to be smaller than the all model. But ridge coefficient estimators only shrink towards zero so it can not do the selection of variables. LASSO can shrink coefficient becomes zero so variable that has a coefficient equal to zero will be selected from the model selected. Based on the results of the analysis LASSO there are only three variables included

TABLE 3
COMPARISON OF COEFFICIENTS COX PH, RIDGE COX AND COXLASSO

Peubah	Cox PH	Ridge	LASSO
x1	-0.0002	0.0001174	.
x2D1	0.36932	1.06E-01	0.228748
x3D1	-0.626387	-1.16E-01	-0.311004
x3D2	-0.076434	6.47E-02	.
x3D3	-0.615765	-7.73E-02	.
x4	0.003127	7.51E-04	0.001890
x5	-0.013975	-2.84E-03	.
x6	-0.08644	-1.87E-02	.
x7	-0.012176	-7.17E-04	.
x8	0.000387	9.60E-05	.
x9	-0.004176	-6.87E-04	.
x10	0.000855	3.09E-04	.

TABLE 6
VALUE RMSEP COMPARISON OF THREE METHODS

Metode	RMSEP
Cox Proportional Hazard	1. 2984
Cox ridge	1. 2274
Cox LASSO	1.5314

in the model, namely X3D1, X2D1 and X4, other variables to zero.

Best model selection is done by calculating the value of RMSEP. Table 4 presents the results RMSEP allegations of patient data at the ICCU RSMY heart patients. RMSEP values obtained regularization Cox ridge smaller than the value of the regularization RMSEP Cox LASSO and Cox Proportional Hazard. This shows that the prediction model using Cox ridge regularization technique is better than the regularization techniques Cox LASSO and Cox Proportional Hazard.

4 CONCLUSION

Data survival cardiac patients in ICCU M. Yunus hospital, observed from January 2011 until December 2011. There were 434 patients with 10 variables was measured for this research. Based on the purpose of this study it can be concluded that in general, patient suffered death in the ICCU mostly on the first 8 days of nursery. The higher the value of variables: Age, Sex, uric acid, creatinine, triglycerides, and LDL, the greater the risk of death. On the other hand, the higher Urea and HDL, the smaller the risk of death. The most influence factors for the survival of patients with cardiovascular disease are gender, status of service, and blood sugar levels. The best model for cardiovascular in ICCU RSMY is cox ridge model.

REFERENCES

[1] Collett, D. 2003. Modelling Survival Data in Medical Research (SecondEdition).London (ENG). Chapman & Hall
[2] Gujarati D. 1992. Basic Econometrics (Translation), 2nd Edition. Interpreting Zeinn S. Jakarta (ID). Erland
[3] Hastie T, Tibshirani R, Friedman J. 2008. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd Ed. New York (US): Springer
[4] Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society Series B 58 (1): 267-288
[5] Tibshirani, R. 1997. The LASSO method for variable selection in the cox models. Statistics inMedicine.16 385-395
[6] World Health Organization, World Health Organization WHO Report 2000, Geneva: WHO, 2001