Comparison of exact, efron and breslow parameter approach method on hazard ratio and stratified cox regression model

Mohamat Fatekurohman¹, Nita Nurmala¹, Dian Anggraeni¹

¹Mathematics of Department, University of Jember, Indonesia

E-mail: mfatekurohman.fmipa@unej.ac.id, nitanurmala8@gmail.com, dian_a.fmipa@unej.ac.id

Abstract. Lungs are the most important organ, in the case of respiratory system. Problems related to disorder of the lungs are various, i.e. pneumonia, emphysema, tuberculosis and lung cancer. Comparing all those problems, lung cancer is the most harmful. Considering about that, the aim of this research applies survival analysis and factors affecting the endurance of the lung cancer patient using comparison of exact, Efron and Breslow parameter approach method on hazard ratio and stratified cox regression model. The data applied are based on the medical records of lung cancer patients in Jember Paru-paru hospital on 2016, east java, Indonesia. The factors affecting the endurance of the lung cancer patients can be classified into several criteria, i.e. sex, age, hemoglobin, leukocytes, erythrocytes, sedimentation rate of blood, therapy status, general condition, body weight. The result shows that exact method of stratified cox regression model is better than other. On the other hand, the endurance of the patients is affected by their age and the general conditions.

1. Introduction

Health is one of the most important thing in our daily. Considering about that, one of the most essential organ to be concerned is lungs. Lungs is the most vital organ, in the case of respiratory system, if lungs get a problem, the respiratory process will get a problem too. That condition brings other problems to other organ, which further end up with mortality. Considering to the fatal effect of lungs disorder, lung cancer is the most harmful since it can contribute to the mortality. Life style, e.g. smoking, is one of the factors increasing the risk of lung cancer. According to National Cancer Institute, a total of 1.685.370 new diagnoses of lung cancer and 589,430 cancer deaths are projected to occur in United States in 2015 [1]. The patients of lung cancer usually survive for short period. The endurance of the patient is affected by many factors in perspective of statistics, the endurance of patient can be predicated using survival analysis.

Survival analysis is a statistical method used to learn the endurance of a patient, which has correlation with time including time origin and end point [2]. Survival function defined as \( S(t) \) is described as the following

\[
S(t) = P \quad (\text{probability of survival with time period})
\]

from definition of cumulative distribution function \( F(t) \) from \( T \), then

\[
S(t) = 1 - P \quad (\text{individual who survives with time period} < t)
\]

\[
= 1 - F(t)
\]

The survival function \( S(t) \) is a nonincreasing function with respect to \( t \) (time). Each of the survival function has the correlation to each other, namely:

hazard function defined in the form of cumulative distribution function \( F(t) \) and density as below

\[
h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{s(t)}
\]

the probability density function defined as the differential of cumulative function
\[
\int_0^t (1 - s(t)) = -s'(t) \tag{3}
\]

\[
-\int H(x)dx = \log S(t) \tag{4}
\]

\[
S(t) = \exp[-H(t)] \tag{5}
\]

we get,

\[
f(t) = h(t) \exp(-H(t)) \tag{6}
\]

according to Collect (1992), hazard proportional model for the unsuccessful survival, denote \(i\) for every time \(t\), can be written as below:

\[
h_i(t) = h_0(t) \exp \left( \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{pi} \right) \tag{7}
\]

The advantage of Cox method is the flexibility of the presence of density function from parametric distribution. Cox method is applied since each independent variable is proportional with respect to time. The proportional assumption of the model can be indicated by plot log \([-\log (s(t))]\) with respect to survival time \(t\) for every time category of descriptive variable \(p\), which form a parallel pattern on different level or uncrossing. The result of plot log \([-\log (s(t))]\) will always increase (exponential function). Furthermore, using plot log \([-\log (s(t))]\) which given information does the variable will proportional in every time [3]. The model log \([-\log (s(t))]\) that is

\[
\log [-\log (s(t))] = \sum_{i=1}^{p} \beta_i x_i + \log [-\log (S_0(t))] \tag{8}
\]

The equation (13) shows that the function of log \([-\log (s(t))]\) is not depended on the time. So, log \([-\log (s(t))]\) function is applicable if it is portrayed toward survival time, then the curve will shape parallel. In the plot of log \([-\log (s(t))]\), the data grouped depend on the level or category on the each variable, then they are needed to be category variable. When the plot of log \([-\log (s(t))]\) shows parallel curve, the proportional hazard assumption is fulfilled. The weakness of log \([-\log (s(t))]\) plot is subjective character, the parallel or not of the curves depend on the method of researcher rate [4]. Hazard ratio value on the regression model cox proportional hazard can be written

\[
\hat{HR} = \frac{h(t|x')}{h(t|x)} \tag{9}
\]

\[
= \frac{h_0(t) \exp \left[ \sum_{i=1}^{p} \beta_i x_i \right]}{h_0(t) \exp \left[ \sum_{i=1}^{p} \beta_i x_i \right]} \tag{10}
\]

\[
= \exp \left[ \sum_{i=1}^{p} \beta_i (x_i - x_i) \right] \tag{11}
\]

The object comparison is equal in the all times or not are depended on the time that indicate proportional hazard assumption are not fulfilled (nonproportional hazard) that is:

\[
\hat{HR} = \frac{h(t|x')}{h(t|x)} \tag{12}
\]

\[
= \frac{h_0(t) \exp \left[ \sum_{i=1}^{p} \beta_i x_i \right]}{h_0(t) \exp \left[ \sum_{i=1}^{p} \beta_i x_i \right]} \tag{13}
\]

\[
= \exp \left[ \sum_{i=1}^{p} \beta_i (x_i - x_i) \right] \tag{14}
\]

the model can be written as:
while the interaction model of stratified cox regression as:

$$h_g(t) = h_0(t) \exp[\beta_{1g}X_1 + \cdots + \beta_{pg}X_p]$$

the difference to the no-interaction model is that the coefficients $\beta$ depend on the strata.

On the survival analysis is possible happened to or more individuals experience occurrence in the same time. Generally, there are 3 methods of parameter approach to complete occurrence in the same time, they are exact, Efron and Breslow parameter approach methods. The estimation equal parameter model as:

**Exact partial likelihood estimator**

$$L \approx \prod_{j=1}^{K} \frac{\exp (S_j \beta)}{\sum_{i: S_i \geq S_j} \exp (S_i \beta)}$$

**Efron partial likelihood estimator**

$$L(\beta) \approx \prod_{i=1}^{r} \frac{\exp (\beta^{*T_i})}{\prod_{j=1}^{d_i} \sum_{k \in R(t_i)} \exp (\beta^{*T_{ik}})}$$

By:

- $d_i$ : the group of all individuals who died at $t_i$
- $x_k$ : the variable of individual that still survive and it is the element of $R(t_i)$ and $d_i$

**Breslow partial likelihood estimator**

$$L(\beta) \approx \prod_{i=1}^{r} \frac{\exp (\beta^{*T_i})}{\sum_{j \in R(t_i)} \exp (\beta^{*T_j})}$$

By:

- $d_i$ : the amount of occurrence in the same time on the beginning of $t_i$
- $R(t_i)$ : the group of individual that has a risk on time $t_i$ which consists of individual who survives on the time $t_i$

In this research, the result of three parameter approaches is compared and modelled by using hazard ratio model and stratified cox regression model, therefore choosing the best model from two model above by looking in the smallest AIC value and using the best parameter approach method.

The method that can be used for choosing the best model is the AIC value. It can be seen in the equation as:

$$AIC = -2 \ln \text{ (maximum likelihood) } + 2m$$

Where $m$ is the number of the estimated parameters and $n$ is the number of observations. The best model is the model which has minimum AIC value [5].

2. Methods

The data of this research are secondary data which are taken from the paru-paru hospital of Jember in 2016 in the form lung cancer patients. There are 64 patients. In this case, the variables used are:

- independent variable ($y$) is the life time (in a day) of lung cancer patient in the hospital of Jember, gender ($x_1$), age ($x_2$), hemoglobin ($x_3$), leukocytes ($x_4$), erythrocytes ($x_5$), sedimentation rate of blood ($x_6$), therapy status ($x_7$), disease history ($x_8$), general ($x_9$), body weight ($x_{10}$).

The AIC (Akaike Information Criteria) value is used to understand the better method for modeling the smallest value. The steps used are:

1. log-rank test
2. proportional test toward ten variables through $\log [-\log S(t)]$
3. comparing the result of parameter estimate with exact, Efron and Breslow method on Hazard Ratio model
4. Comparing the result of parameter estimate with exact, Efron and Breslow method on Stratified Cox Regression model
5. Choosing the best model and the best estimate approach by looking in the smallest AIC value

3. Result
The occurrence which is analysed in this research is died occurrence when the patients are in the Jember paru-paru hospital. For the patients who are out from another hospital then include in the censored data. There are three types of censor in this research. The patient who includes in the research in the different time of January to May 2016, by “1” is the patient who are not censored and “0” is the patients who are censored. There are 10 variables for shaping Hazard ratio and stratified cox regression modal from lung cancer data, namely gender, age, haemoglobin, leukocytes, erythrocytes, sedimentation rate of blood, therapy status, disease history, general condition and body weight. The result of R and SAS program for every variable which is analysed is

Table 1. Log-rank test for ten variables

| Variable                      | Test statistics | \((X^2)_{table}\) | significant | decision |
|-------------------------------|-----------------|----------------------|-------------|----------|
| Gender                        | 2.9             | 3.84                 | 0.09        | Accept \(H_0\) |
| Age                           | 14              | 7.82                 | 0           | reject \(H_0\)  |
| Hemoglobin                    | 0.1             | 5.99                 | 0.96        | Accept \(H_0\)  |
| Leukocytes                    | 0.6             | 5.99                 | 0.75        | Accept \(H_0\)  |
| Erythrocytes                  | 1.3             | 5.99                 | 0.53        | Accept \(H_0\)  |
| Sedimentation rate of blood   | 3.3             | 3.84                 | 0.07        | Accept \(H_0\)  |
| Therapy status                | 1.4             | 3.84                 | 0.24        | Accept \(H_0\)  |
| Disease history               | 0.2             | 3.41                 | 0.68        | Accept \(H_0\)  |
| General condition             | 6               | 5.99                 | 0.05        | reject \(H_0\)  |
| Body weight                   | 5.1             | 9.49                 | 0.28        | Accept \(H_0\)  |

Testing the assumption of proportional hazards modeling to ten explanatory variables through the plot 
\(log [−log S (t)]\) obtained 10 variables is
Figure 1. Plot against time $\log \left[ -\log (S(t)) \right]$
The results of the comparison exact, Efron and Breslow parameter approach method on a model hazard ratio from the combination of 10 variables got the best models from each of the parameter estimation approach are:

- hazard ratio model using exact parameter approach of AIC = 178.938
  \[
  h(t, X) = h_0(t) \exp \left[ -0.15503x_1g_1(t) + 0.04754x_5g_5(t) - 0.28071x_6g_6(t) + 0.2951x_7g_7(t) - 0.31782x_9g_9(t) \right]
  \]

- hazard ratio model using Efron parameter approach AIC = 307.86
  \[
  h(t, X) = h_0(t) \exp \left[ -0.10907x_1g_1(t) + 0.02795x_5g_5(t) - 0.27743x_6g_6(t) + 0.29232x_7g_7(t) - 0.27633x_9g_9(t) \right]
  \]

- hazard ratio model using Breslow parameter approach with AIC = 325.86
  \[
  h(t, X) = h_0(t) \exp \left[ -0.09034x_1g_1(t) + 0.02608x_5g_5(t) - 0.23051x_6g_6(t) + 0.26914x_7g_7(t) - 0.2467x_9g_9(t) \right]
  \]

The results of the comparison exact, Efron and Breslow parameter approach method on a stratified Cox regression models are:

- stratified Cox regression model using exact parameter approach of AIC = 99.091
  \[
  h_g(t) = h_{0g}(t) \exp \left[ -0.17829x_1g_1(t) + 0.0904x_5g_5(t) - 0.27545x_6g_6(t) + 0.20410x_7g_7(t) - 0.48958x_9g_9(t) \right]
  \]

  by stating strata, \( g = 1, 2, \ldots, 12 \)

- stratified Cox regression model using Efron parameter approach AIC = 141.94
  \[
  h_g(t) = h_{0g}(t) \exp \left[ -0.06363x_1g_1(t) + 0.07273x_5g_5(t) - 0.26984x_6g_6(t) + 0.22026x_7g_7(t) - 0.36849g_9(t) \right]
  \]

  by stating strata, \( g = 1, 2, \ldots, 12 \)

- stratified Cox regression model using Breslow parameter approach with AIC = 157.698
  \[
  h_g(t) = h_{0g}(t) \exp \left[ -0.05366x_1g_1(t) + 0.06394x_5g_5(t) - 0.25124x_6g_6(t) + 0.18644x_7g_7(t) - 0.32879g_9(t) \right]
  \]

  by stating strata, \( g = 1, 2, \ldots, 12 \)

Of the three approaches parameter estimation showed that exact method is the best parameter estimation approach for modeling the hazard ratio and stratified Cox regression model with the best are:

- hazard ratio model with the exact parameter approach of AIC = 178.938
  \[
  h(t, X) = h_0(t) \exp \left[ -0.15503x_1g_1(t) + 0.04754x_5g_5(t) - 0.28071x_6g_6(t) + 0.2951x_7g_7(t) - 0.31782x_9g_9(t) \right]
  \]

- stratified Cox regression model with the exact parameter approach of AIC = 99.091
  \[
  h_g(t) = h_{0g}(t) \exp \left[ -0.17829x_1g_1(t) + 0.0904x_5g_5(t) - 0.27545x_6g_6(t) + 0.20410x_7g_7(t) - 0.48958x_9g_9(t) \right]
  \]

  by stating strata, \( g = 1, 2, \ldots, 12 \)

4. Discussion

Log-rank test is often used to see the difference of life endurance in the group with related group. From the log-rank test showed that:

1. gender, HB, leukocytes, erythrocytes, LED, therapy status, history of disease and body weight does not cause any difference in the function of survival in patients with lung cancer
2. differences in age and general condition causes the difference in the function of survival in patients with lung cancer

According testing the assumption of proportional hazards modelling to ten explanatory variables through the plot log [-log S (t)] obtained 10 variables are not proportional for forming a graph that is not parallel. Nonproportional hazard because the study time in the model it uses a hazard ratio and stratified cox regression model. Then from the best model hazard ratio and stratified cox regression...
model to estimate the exact parameter approach, obtained the exact parameter estimation approach will provide the best model with the smallest AIC by using stratified Cox regression model.

5. Conclusion
Best parameter estimation approach is to use the exact method by using a stratified Cox regression model, namely
\[ h_g(t) = h_0g(t) \exp \left[ -0.17829x_1g_1(t) + 0.0904x_5g_5(t) - 0.27545x_6g_6(t) + 0.20410x_7g_7(t) - 0.48958x_9g_9(t) \right] \]
by stating strata, \( g = 1, 2, \ldots, 12 \)

Acknowledgement
Thanks to the Director General of Research and Community Service of the Ministry of Research and Technology Higher Education for funding support through Post Graduate Doctoral Research 2017.

References
[1] Siegel R L, Miller K D and Jemal A 2015 Cancer Statistics 65 5-29
[2] Kleinbaum D G & Klein M 2011 Survival Analysis a Self-Learning Text Third Edition (New York : Springer)
[3] Ata N and Tekin M T 2007 Cox Regression Model with Nonproportional Hazard Applied to Lung Cancer Survival Data Hacettepe Journal of Mathematics and Statistics 2 157-167
[4] Collect D 2003 Modelling Survival Data in Medical Research (London : Chapman and Hall)
[5] Wang Yand Liu Q 2006 Comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in Selection of Stock-recruitment Relationships Fisheries Research of China Ocean University 77 220-225