Analysis of survival in breast cancer patients by using different parametric models

Syahila Enera Amran, M Asrul Afendi Abdullah, Kek Sie Long and Siti Afiqah Muhamad Jamil

Mathematics and Statistics Department, Universiti Tun Hussein Onn, Jln Parit Raja, 86400, Batu Pahat, Johor, Malaysia.

Email: syahilaamran@gmail.com, afendi@uthm.edu.my

Abstract. In biomedical applications or clinical trials, right censoring was often arising when studying the time to event data. In this case, some individuals are still alive at the end of the study or lost to follow up at a certain time. It is an important issue to handle the censoring data in order to prevent any bias information in the analysis. Therefore, this study was carried out to analyze the right censoring data with three different parametric models: exponential model, Weibull model and log-logistic models. Data of breast cancer patients from Hospital Sultan Ismail, Johor Bahru from 30 December 2008 until 15 February 2017 was used in this study to illustrate the right censoring data. Besides, the covariates included in this study are the time of breast cancer infection patients survive $t$, age of each patients $X_1$ and treatment given to the patients $X_2$. In order to determine the best parametric models in analysing survival of breast cancer patients, the performance of each model was compare based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log-likelihood value using statistical software R. When analysing the breast cancer data, all three distributions were shown consistency of data with the line graph of cumulative hazard function resembles a straight line going through the origin. As the result, log-logistic model was the best fitted parametric model compared with exponential and Weibull model since it has the smallest value in AIC and BIC, also the biggest value in log-likelihood.

1. Introduction

Survival analysis was the phrase that is used to describe the analysis of data where the outcome variable was the time until the event of interest occurred. The event can be either death, occurrence of the disease, divorce, marriage or any designated experience of interest that may happen to an individual. Unfortunately, survival times with censoring were frequently occurred in many medical and reliability studies (Huang, J., 2007). The survival time of an individual is said to be censored when there have some information about the individual but the exact survival time is unknown. Like Guo, S. & Zeng, D. (2014) said, only the partial information of survival time was observed instead of the accurate information.

There were three types of censoring data; right censoring, left censoring and interval censoring. When the event time lies on an interval of the form $(L, R)$, where $L$ is the last time the patient seen without having the disease, and $R$ is the first time the patient appeared with the disease, the event was said to be the interval censoring. Therefore, when the time of event occurs after the
study end, \( R = \infty \), the event was right censoring while the left censoring occurs when the event was happened before the study started with \( L = 0 \) (Ma, L., 2014).

Analyzing censoring data either right censored, left censored or interval censored data has been a major challenge in medical research. Even though many non- and semi-parametric statistical methods were developed, these methods were not well understood by the medical community due to their complexity. Parametric methods were easy to implement, but unfortunately this method received little attention in practice because the impact of model misspecification on the estimation results were not well understood (Gong, Q. & Fang, L., 2013). Therefore, exponential distribution, Weibull distribution and log-logistic distribution was used to analyze the right censoring data.

Breast cancer was the most common cancer among women in Malaysia. About 1 out of 20 chance of Malaysian women were diagnosed the breast cancer in lifetime (J Leong, B et al, 2007). Breast cancer treatment can be divided into two types; local therapy and systemic therapy (Yip, C. et al). Local therapy includes surgery and radiation therapy for the breast cancer are while systematic therapy includes chemotherapy, hormone therapy and targeted therapy. According to World Health Organisation (WHO), 560 thousand women will die caused by the breast cancer in 2015 and new cases were diagnosed an estimated 1.6 million. It shows that the incidence rate of breast cancer continues to show an increasing trend worldwide. Hence, this study will be analyze by using the breast cancer data to improve the survival rate of the breast cancer incidence. In order to determine the best parametric model in breast cancer patients, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log-likelihood value were calculated and compared.

2. Material and methods

This study was on comparing the performance of parametric model in right censored data by using exponential distribution, Weibull distribution and log-logistic distribution. By using different parametric models, data of breast cancer patients from Hospital Sultan Ismail, Johor Bahru was used in this study. The durations of breast cancer infection patients survive, age of each patients and the treatment given to the patients will be analyzed as the covariates. The data was involved with 38 of female infection patients of whom 8 of patients were treated by local therapy, 21 of them receive systematic therapy and the other 9 patients were treated by local and systematic therapy. Statistical software R was used to find out which of the models can give the best fitted in right censored data of breast cancer patients. As discussed in Prentice, R. L., & Gloeckler, L. A. (1978), all the three parametric distributions are analyze and compare the performance by using AIC, BIC and log-likelihood value.

2.1. Exponential distribution

The exponential distribution is the simplest and most important distribution in survival studies. It describes the arrival time of a randomly recurring independent event sequence. The exponential distribution can serve as the baseline for more complex models, given its constant failure rate

\[ h(t) = \lambda \quad \text{where } t \geq 0, \quad \lambda > 0 \quad (1) \]

The large \( \lambda \) indicates high risk and short survival while a small \( \lambda \) indicates low risk and long survival. The probability density function \( f(t) \) and survivor function \( S(t) \) are found as the following form:

\[ f(t) = \lambda e^{-\lambda t} \quad (2) \]

\[ S(t) = e^{-\lambda t} \quad (3) \]

When \( \theta = \lambda = 1 \), it is termed standard exponential distribution. In addition, exponential distribution is special cases of both the Weibull and gamma distributions.
2.2. Weibull distribution

Weibull distribution is the most commonly used distribution in reliability when data is censored since it has the flexibility of the scale and shape parameters. According to Waloddi Weibull who introduced the Weibull distribution in 1937, the advantage of Weibull is the ability to provide reasonably accurate failure analysis and failure forecasts with extremely small samples. The Weibull distribution has a hazard function of the following form:

\[ h(t) = \lambda \beta (\lambda t)^{\beta-1} \]  

where scale parameter \( \lambda > 0 \) and the shape parameter \( \beta > 0 \). The value of \( \beta \) determines the shape of the distribution curve and the value of \( \lambda \) determines its scaling. It includes the exponential distribution as the special case where \( \beta = 1 \). The probability density function \( f(t) \) and survivor function \( S(t) \) of the distribution are given by:

\[ f(t) = \lambda \beta (\lambda t)^{\beta-1} e^{-(\lambda t)^\beta} \]

\[ S(t) = e^{-(\lambda t)^\beta} \]

2.3. Log-logistic distribution

Log-logistic distribution has a hazard rate with hump-shaped, which is increases initially then, decreases. It has a survival function and hazard rate that has a closed form expression, as contrasted with the log normal distribution which also has a hump-shaped hazard rate. The log-logistic distribution has the hazard function \( h(t) \) as of the following form:

\[ h(t) = \left( \frac{\beta}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1} \left[ 1 + \left( \frac{t}{\alpha} \right)^{\beta} \right]^{-2} \]

where \( \alpha > 0 \) is the scale parameter and \( \beta > 0 \) is the shape parameter. The probability density function \( f(t) \) and the survivor function \( S(t) \) are, respectively:

\[ f(t) = \left( \frac{\beta}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1} \left[ 1 + \left( \frac{t}{\alpha} \right)^{\beta} \right]^{-2} \]

\[ S(t) = \left[ 1 + \left( \frac{t}{\alpha} \right)^{\beta} \right]^{-1} \]

2.4. Parametric error measurements

In order to determine the best parametric model in partly interval-censored data, the mean square error (MSE), mean bias (MB), Akaike information criterion (AIC) and Bayesian information criterion (BIC) will be calculated.

2.4.1. Akaike information criterion

Akaike Information Criterion (AIC) is a measure of selecting a model from a set of models. AIC estimates the quality of each model, relative to each of the other models. The AIC is given by:

\[ AIC = -2 \times \log(\text{likelihood}) + 2(k) \]  

where \( k \) is the number of parameters in model. Thus, \( k = 1 \) for the exponential model and \( k = 2 \) for the Weibull and log-logistic models. AIC can also be calculated using residual sums of squares from regression:

\[ AIC = n \times \log\left(\frac{\text{RSS}}{n}\right) + 2(k) \]

\[ AIC = -2 \times \log(\text{likelihood}) + 2(k) \]
where \( n \) is the number of data points (observations) and RSS is the residual sums of squares. Smaller AIC indicate a better model fit.

2.4.2. Bayesian information criterion
The Bayesian information criterion (BIC) or also known as Schwarz criterion (SBC, SBIC) is a criterion for model selection among a finite set of models. It is based on the likelihood function and hence, it is closely related to the Akaike information criterion (AIC). The BIC is given by:

\[
BIC = -2 \log(\text{likelihood}) + k \log(n)
\]

(12)

where \( n \) is the sample size and \( k \) is the number of covariates including an intercept. Same as AIC, BIC also has the value of \( k=1 \) is for the exponential model and \( k=2 \) for the Weibull and log-logistic models. Smaller BIC indicate a better model fit. The BIC generally penalizes free parameters more strongly than does the Akaike information criterion, though it depends on the size of \( n \) and relative magnitude of \( n \) and \( k \).

2.4.3. Log-likelihood value
The estimated parameters of three parametric models were estimated by using maximum likelihood functions as discussed before. The selection of the best fit depends on the likelihood values of the observe data under three parametric model. The log-likelihood function is given by:

\[
\log L(\theta; y) = \sum_{i=1}^{n} \log f_i(y_i; \theta)
\]

(13)

The model that gives and the highest likelihood value give the best fit model.

3. Result and discussion
The survival time of breast cancer patients at Hospital Sultan Ismail had an average of 32 months. Out of 38 female patients, 33 of them are censored or still alive until the end time of study while the other 5 patients died during the study. Besides that, the average age of women diagnosed with the breast cancer is in the range of 52 years old. The youngest woman is diagnosed at the age of 26 years old while the oldest women infected with the disease is in the range of 60 years. Based on the Kaplan-Meier graph, age group between 58 to 73 years has the highest survival probability in breast cancer patients followed by the age group between 48 to 57 years. Meanwhile, age group 26 to 47 years has the lowest survival probability.

3.1. Chi-square test for independence
The chi-square test for independence is used to compare between two sets of data or categorical variables in order to see whether the distributions of categorical variables differ from each another. This study perform a chi-square to see whether the age of breast cancer patients and treatments given have a relationship with the breast cancer survival. The null of the chi-squared test is the two variables are independent and the alternate hypothesis is that they are dependent with each other. As a results, the treatment given to the breast cancer patients is more related with the patient’s status compared with the age of the patients since the p-value of variable treatment is less than 0.05.

3.2. Model results
When analyze the parametric models, cumulative hazard plots are used for examining the distributional model assumptions visually for survival analysis. The cumulative hazard plot of the exponential, Weibull and log-logistic distributions shows a reasonable straight-line fit through the origin. Hence, all the three models are fitted and shows consistency of the data.
Table 1. Parametric Model Fitted to Breast Cancer Data.

| S. No. | Covariates | Exponential | Weibull | Log-logistic |
|--------|------------|-------------|---------|--------------|
|        |            | Coef.       | S. E.   | Coef.        | S. E.   | Coef.    | S. E.   |
| 1      | Age        | -0.05800    | 0.04370 | -0.05352     | 0.01920 | -0.06722 | 0.01380 |
| 2      | Treatment  | -17.51500   | 1320.0000 | -8.00361     | 359.0000 | -6.20245 | 26.40000 |

Table 2. Parametric Model Selection in Breast Cancer Data.

| Distributions | Measurements |
|---------------|--------------|
|               | AIC          | BIC          | Log-likelihood Value |
| Exponential   | 48.80000     | 48.37978     | -23.40000             |
| Weibull       | 46.40000     | 45.55957     | -21.20000             |
| Log-logistic  | 44.00000     | 43.15957     | -20.00000             |

The parametric models were fitted using `survreg` package in statistical software R and the results are presented in Table 1. From the table we see that the two covariates namely age and treatment are significantly associated with the survival time under all the model assumptions. Among the models, the log-logistic has the lowest value of standard error compared to all other models. The parametric model selection of each distribution using AIC, BIC and log-likelihood value are given in Table 2. When comparing the parametric models, log-logistic distribution had the smallest value in AIC and BIC while for the log-likelihood value, it has the biggest value. It is proving that log-logistic distribution is the best parametric model for the breast cancer patient’s data.

4. Conclusion

The aim of this study is to determine the best parametric model when using breast cancer patients data as the right censoring. Three different parametric are involved in this study namely exponential model, Weibull model and log-logistic model. In order to determine the best parametric model, three different measurements of model selection are calculated by using statistical software R.

About 38 female of breast cancer infected are involved in this study and their time of survival, age when diagnosed and the treatment given are recorded as the covariates. Based on the data collection, the minimum time of survival is only one month while the maximum survival time of breast cancer infection is about 11 years. By using Kaplan-meier curve, it shows that age group between 58 to 73 years has the highest survival probability and the lowest survival probability is the women age between 26 to 47 years old. Besides that, the age of patients is independent with the status of breast cancer patients while the treatment given shows a dependent variable to the patient’s status based on the chi-square test.

When analyzing the breast cancer data, all the three distributions are shown consistency of data with the line graph of cumulative hazard function resembles a straight line going through the origin. Therefore, exponential model, Weibull model and log-logistic models are appropriate to use for the survival analysis in breast cancer.

Besides, the measurements of model selection are calculated using AIC, BIC and log-likelihood value. When comparing the three distributions, log-logistic distribution has the smallest
value in AIC and BIC while it has the biggest value in log-likelihood compared with exponential and Weibull distribution. Hence, the log-logistic distribution is the best fitted parametric model for the breast cancer patients data. However, Vallinayagam, V. et al (2014) and Pourhoseingholi, M. A et al (2007) find out that lognormal was the best fitted among parametric models when using AIC as the criterion of evaluation. As the future research, this analysis should be include others parametric models such as gamma and lognormal distributions.

Acknowledgement
I would like to acknowledge Universiti Tun Hussein Onn Malaysia, UTHM for supporting this research project.

5. References
[1] Yau.K.K.W and Lee.A.H. (2001). Zero inflated poisson regression with random error to evaluate an occupational injury prevention program. Statistic in Medicine, 20: 2907{2920.
[2] Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. Statistica Sinica, pp. 501–519.
[3] Guo, S. & Zeng, D. (2014). An overview of semiparametric models in survival analysis. Journal of Statistical Planning and Inference, 151, pp. 1–16.
[4] Ma, L., Feng, Y., Chen, D. G. D., & Sun, J. (2014). Interval-Censored Time-to-Event Data and Their Applications in Clinical Trials. Clinical Trial Biostatistics and Biopharmaceutical Applications, 307.
[5] Gong, Q. & Fang, L. (2013). Comparison of different parametric proportional hazards models for interval-censored data: A simulation study. Contemporary clinical trials, 36(1), pp. 276–283.
[6] J Leong, B., Chuah, J.A., Kumar, V.M. & Yip, C.H. (2007). Breast cancer in Sabah, Malaysia: a two year prospective study. Asian Pac J Cancer Prev, 8(4), pp. 525–9.
[7] Yip, C., Bhoo, P.N. & Teo, S. (2014). A review of breast cancer research in malaysia. Med J Malaysia, 69, pp. 8–12.
[8] Prentice, R. L., & Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. Biometrics, 57-67.
[9] Vallinayagam, V., Prathap, S., & Venkatesan, P. (2014). Parametric Regression Models in the Analysis of Breast Cancer Survival Data. Int J Sci Tech, 3, 163-7.
[10] Pourhoseingholi, M. A., Hajizadeh, E., Moghimi Dehkordi, B., Safaee, A., Abadi, A., & Zali, M. R. (2007). Comparing Cox regression and parametric models for survival patients with gastric carcinoma. Asian Pacific Journal of Cancer Prevention, 8(3), 412.