The Log Beta Generalized Weibull Regression Model for Lifetime Data

N Chanasriphum, P Seenoi and W Srisodaphol

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand

Email: palakorns@kku.ac.th

Abstract. A log beta generalized Weibull regression model was proposed by modification from beta generalized Weibull distribution. To predict the survival time, the maximum likelihood method was used to estimate parameters in the proposed regression model and then, the proposed regression model was applied with lifetime data. The Akaike information criterion, Bayesian information criterion and likelihood ratio statistic are criterion to compare the performance of the proposed regression model with sub models which are log beta Weibull regression model, log generalized Weibull regression and log Weibull regression model. The result showed that the proposed regression model performed better than the sub models.

1. Introduction

Lifetime data is lifetime measurement data such as the duration of the equipment is repaired or expired in a continuous period, duration of disease, duration of cancer cell growth, time of rash disease and survival time of the patient etc. In the field of medical and engineering, the data are measured in form of survival time and number of failures for the components. Survival analysis is an analyzing the expected duration of time until one or more events happen (time to event). An analyzing with time data is help you to get the details of the data rather than cutting them out such as when the doctor follows up patient’s death for any diseases, they found that patient is die a lot in the first phase of study and the death decreased in later. In this case if time to die of each patient is not collected, we will know only the proportion of patient’s death at the end of study. Survival analysis with time data makes the results more efficient, which can be applied in several of study. The example of survival analysis such as in the engineering field, survival analysis is used to study the lifetime of the machines, in the medical field, survival analysis is used to follow up patient’s disease, patient’s death or well.

The lifetime distribution is a function showing the lifetime data which was used to describe the probability distribution in the form of hazard rate such as engineering field used in reliability, medical field used in survival. The probability distributions that were applied to the lifetime data such as the exponential distribution, gamma distribution which was proposed by Euler in 1783 [1], log normal distribution which
was proposed by McAlister and Galton in 1879 and Weibull distribution which was proposed by Waloddi Weibull in 1937 [2]. Moreover, there are many researchers have proposed the distribution that relate with lifetime data analysis, such as generalized Weibull distribution which was proposed by Hutson and Mudholkar in 1996 [3], [4] proposed beta Weibull distribution and beta generalized Weibull distribution [5].

In addition, the lifetime data can also use for regression analysis, which is the statistical methods for the analysis of relationships between two variables which is dependent variable or response variable and independent variables or predictor variables. The values of independent variables are known. The objective of regression analysis is to create a regression model which has the ability to predict the value of a dependent variable based on values of independent variables. If the dependent variable and independent variable has a non-linear relationship, called a non-linear regression model, we have to convert the data to make the relation between the dependent variables and independent variables to a linear before analyze. In 1979 and 1982, [6] presented the non-linear regression model called the log Weibull regression model, which is developed from Weibull distribution to predict the survival chances in form of the survival function. Moreover, there are many researchers used different distributions to create a regression model to predict the survival chances such as log Burr XII regression model [7], exponentiated Weibull regression model or log generalized Weibull regression model [8], log beta Weibull regression model [9] and log beta log logistic regression model [10].

In this paper, we will create a log beta generalized Weibull regression model from the probability density function of the beta generalized Weibull distribution [5] which are appropriate for lifetime data. This model is created for predicting the survival chance of lifetime data when have covariate variables. The performance of new regression models is compared with log generalized Weibull regression model, log beta Weibull regression model and log Weibull regression model by using criterion AIC, BIC and LR statistic to measures of regression model.

2. The log beta generalized Weibull Distribution

The beta generalized Weibull distribution (BGW) [5] is the distribution with five parameters, \(a > 0, b > 0, \alpha > 0, \beta > 0\) and \(\lambda > 0\). Let \(T\) is lifetime data which has a probability density function as

\[
f(t; a, b, \alpha, \beta, \lambda) = \frac{\alpha \beta \lambda ^{\beta - 1} t^{\beta - 1}}{B(a, b)} \left[1 - \exp\left(-\lambda t^\beta\right)\right]^{\alpha - 1} \left[1 - \left(1 - \exp\left(-\lambda t^\beta\right)\right)^{\alpha}\right]^{\gamma - 1} \exp\left(-\lambda t^\beta\right), \quad t > 0
\]

where \(B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a + b)}\) is the beta function and \(\Gamma(\cdot)\) is the gamma function. Here, \(a, b, \alpha\) and \(\beta\) are the shape parameters, and \(\lambda\) is a scale parameter. The survival function \(S(t)\) and hazard function \(h(t)\) corresponding to a random variable \(T\) for BGW distribution are given respectively by

\[
S(t) = 1 - \int_{[1-\exp(-\lambda t^\beta)]^{\gamma}}^{(a,b)}\, (a,b)
\]

and

\[
h(t) = \frac{a \beta \lambda ^{\beta - 1} (1 - \exp(-\lambda t^\beta))^{\alpha - 1}}{B(a, b)} \left\{1 - \left[1 - \exp(-\lambda t^\beta)\right]^{\alpha}\right\}^{\gamma - 1} \exp(-\lambda t^\beta)
\]

\]
where \( I_y(a,b) = B(a,b) \int_0^1 w^{a-1}(1-w)^{b-1} dw \) is the incomplete beta function ratio.

Let \( T \) be a random variable having the BGW distribution in equation (1). Denote that the random variable \( Y = \log(T) \), we reparametrize in terms of \( \sigma = \beta^{-1} \) and \( \mu = \log(1/\lambda) \). The probability density function of \( Y \) can be expressed as

\[
f(y; a, b, \alpha, \mu, \sigma) = \frac{\alpha}{\sigma B(a, b)} \exp\left(\frac{y - \mu}{\sigma}\right) \exp\left(-\frac{y - \mu}{\sigma}\right) \left[1 - \exp\left(-\frac{y - \mu}{\sigma}\right)\right]^{\alpha-1} \\
\times \left[1 - \left(1 - \exp\left(-\frac{y - \mu}{\sigma}\right)\right)^\alpha\right]^{b-1}
\]

where \(-\infty < y < \infty\), \(-\infty < \mu < \infty\), \(a > 0\), \(b > 0\), \(\alpha > 0\) and \(\sigma > 0\). So, \( Y \) has the log beta generalized Weibull distribution (LBGW) or \( Y \sim \text{LBGW}(a, b, \alpha, \mu, \sigma) \), where \( \mu \) is a location parameter, \( \sigma \) is scale parameter and \( a \), \( b \) and \( \alpha \) are shape parameters.

The cumulative distribution function of \( Y \) can be expressed as

\[
F(y; a, b, \alpha, \mu, \sigma) = \frac{\alpha}{\sigma B(a, b)} \left[1 - \exp\left(-\frac{y - \mu}{\sigma}\right)\right]^{\alpha-1} (a, b).
\]

The survival function of \( Y \) takes the form

\[
S(y; a, b, \alpha, \mu, \sigma) = 1 - F(y; a, b, \alpha, \mu, \sigma) = 1 - \frac{\alpha}{\sigma B(a, b)} \left[1 - \exp\left(-\frac{y - \mu}{\sigma}\right)\right]^{\alpha-1} (a, b).
\]

Further, we define the standard random variable \( Z = (Y - \mu) / \sigma \) with probability density function

\[
f(z; a, b, \alpha) = \frac{\alpha}{B(a, b)} \exp\{z - \exp(z)\} \{1 - \exp(-\exp(z))\}^{\alpha-1} \{1 - (1 - \exp(-\exp(z)))\}^{b-1}
\]

where \(-\infty < z < \infty\).

### 3. The log beta generalized Weibull regression model

Based on the LBGW probability density function, we propose a log linear regression model corresponding the dependent variable \( y_i \) and the independent variable vector as follows

\[
y_i = x_i \beta + \sigma z_i \quad ; i = 1, \ldots, n
\]
where the random error \( z_i \) has probability density function in equation (6), \( y_i = \log(t_i) , a > 0 , b > 0 , \alpha > 0 , \sigma > 0 \), \( \mu_i = x_i \beta , \ x_i = (1, x_{i1}, \ldots, x_{ik}) \) and \( \beta = (\beta_0, \beta_1, \ldots, \beta_k)' \).

Let \( Y_1, \ldots, Y_n \) be independent and identical distribution random sample for parameters \( \theta = \{a, b, \alpha, \beta', \sigma\} \) with probability density function \( f(y | \theta) \). The likelihood function is

\[
L(\theta | y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i | \theta).
\]

The maximum likelihood estimates of \( \theta \) are

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} L(\theta | y_1, \ldots, y_n).
\]

We use the maximum likelihood estimates in equation (10) and their invariance property to predict the survival time for \( t = \exp(y) \), where \( \hat{\beta} = 1/\hat{\sigma} \) and \( \hat{\lambda} = 1/\exp(x\hat{\beta}) \) by using the survival function in equation (11),

\[
S(t; \hat{\alpha}, \hat{\beta}, \hat{\lambda}) = 1 - I_{[\exp(-(\hat{\lambda}y)^\alpha)^\beta]}(\hat{\alpha}, \hat{\beta}).
\]

To measure performance of regression models, the criterion which are used in this study are Akaike information criterion (AIC) and Bayesian information criterion (BIC) [12]. The AIC and BIC are respectively showed as

\[
AIC = -2\log(\text{likelihood}) + 2(p),
\]

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

where \( p \) is the number of free parameters in the model and \( n \) is sample size.

Also, we use the likelihood ratio (LR) statistic for testing the goodness of fit of LBGW regression model comparing with sub models; LBW, LGW and LW regression models. The hypotheses are tested as follows.

(i) \( H_0: \alpha = 1 \) versus \( H_1: H_0 \) is not true (If \( H_0 \) is rejected, LBGW regression model will be more appropriate than LBW regression model.)

(ii) \( H_0: a, b = 1 \) versus \( H_1: H_0 \) is not true (If \( H_0 \) is rejected, LBGW regression model will be more appropriate than LGW regression model.)

(iii) \( H_0: a, b, \alpha = 1 \) versus \( H_1: H_0 \) is not true (If \( H_0 \) is rejected, LBGW regression model will be more appropriate than LW regression model.)

This statistic is given by [5]

\[
D = -2\log\left( \frac{\text{likelihood under } H_0}{\text{likelihood under } H_1} \right)
\]
4. Application

In this section, an application of result will be provided by real data. We use the data set of liver patients of Mayo Clinic Primary Biliary Cirrhosis (PBC) which conducted between 1974 and 1984 [11]. A total 407 patients are used for the fit LBGW regression model. The data consist of random response variable which is the number of days until each patient died ($t_i$), when $y_i = \log(t_i)$. The following explanatory variables were associated with each patient (for $i=1,\ldots,407$):

- $x_{i1}$ is Age (years)
- $x_{i2}$ is Platelet Count
- $x_{i3}$ is serum albumin (g/dl)

Then, we present the results of fitting the model in following form,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma z_i$$

where $i=1,\ldots,407$.

We use AIC, BIC and LR statistic to measure performance of regression models. The package nlm in R is use for calculating MLEs $\hat{\theta}$. The relation of parameters among LBGW regression model and sub models, is showed in figure 1.

![Diagram of sub models of the log beta generalized Weibull regression model.](image)

**Figure 1.** Sub models of the log beta generalized Weibull regression model.

From figure 1, LBGW with $\alpha=1$ become LBW regression model. LBW with $a,b=1$ become LW regression model. Moreover, LBGW with $a,b=1$ become LGW regression model. LGW with $\alpha=1$ become LW regression model.

The estimated values of parameters for LBGW, LBW, LGW and LW regression models fitting with the real data are reported in table 1. Moreover, we use the survival time of $T$ in equation (11) to predict survival time of PBC patients at $t$ days. Plots of survival time for PBC patients based on 4 regression models are showed in figure 2.
Table 1. The MLEs of the parameters for LBGW, LBW, LGW and LW regression models fitting to the real data.

| Model | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{\beta}_a$ | $\hat{\beta}_b$ | $\hat{\beta}_c$ | $\hat{\sigma}$ |
|-------|-----------|-----------|-----------|-----------------|-----------------|-----------------|----------------|
| LBGW  | 7.8147    | 0.2201    | 0.0671    | 5.1412          | -0.0016         | 0.0005          | 0.4999         |
| LBW   | 1.0097    | 190.8108  | -         | 8.6004          | -0.0025         | 0.0006          | 0.5660         |
| LGW   | -         | -         | 1.2041    | 5.0197          | 0.0006          | 0.0007          | 0.6582         |
| LW    | -         | -         | -         | 7.0249          | -0.0089         | 0.0003          | 0.2837         |

Figure 2. Estimated survival time for PBC patients versus the Kaplan-Meier curves.

From figure 2, we obtain that the survival time of 4 regression models have the similarity result that is the line of survival time approached the Kaplan-Meier curves. Therefore, we have to report the values of AIC, BIC and LR statistic of LBGW, LBW, LGW and LW regression models in table 2-3, respectively for finding a clearly result.
From table 2, the AIC and BIC values of LBGW regression model are smallest. Therefore, we conclude that LBGW regression model is best fit when comparing with sub models. Moreover, the p-value in table 3 indicates that the null hypothesis is rejected for all cases. Therefore, we conclude that LBGW regression model fits the real data more than sub models at significant level of 5%.

Table 3. Values of LR statistic and p-value for comparison models.

| Hypothesis            | Values of LR Statistic | p-value |
|-----------------------|------------------------|---------|
| LBW versus LBGW       | 8.8922                 | 0.0029  |
| LGW versus LBGW       | 13.8230                | 0.0002  |
| LW versus LBGW        | 26.1490                | < 0.0000|

5. Conclusions
In this study, we proposed a log beta generalized Weibull regression model which is suitable for lifetime data. We estimated parameters by using maximum likelihood estimation and compared the proposed regression model with sub models; LBW regression model, LGW regression model and LW regression model by using the AIC, BIC criterions and likelihood ratio statistic. Moreover, we applied these regression models with real data to predict survival time. Finally, we found that the proposed regression model performs better than sub models.

Acknowledgements
The authors are very thankful to Department of Statistics, Faculty of Science and Graduate School, Khon Kaen University for financial support.

References
[1] Sebah P and Gourdon X 2002 Introduction to the Gamma Function. [cited 2019 Jan 25]. Available from: https://www.csie.ntu.edu.tw/~b89089/link/gammaFunction.pdf
[2] Almalki S J and Nadarajah 2014 Reliability Engineering and System Safety. 124 32-55
[3] Lai C D 2014 Generalized Weibull Distributions Springer Briefs in Statistics (New York: Springer-Verlag Berlin Heidelberg) p 23-75
[4] Famoye F, Lee C and Olumilade O Statistical Theory and Application. 4(2) 121-136
[5] Singla N, Jain K, and Sharma S K 2012 Reliability Engineering and System Safety. 5-15
[6] Roger W H and Hanley J M 1982 The Rand Corporation. 586-596
[7] Silva G O, Ortega E M M, Cancho V G and Barreto M L 2008 *Computational Statistics & Data Analysis*. 52(7) 3820-3842
[8] Hashimoto E M, Ortega E M M, Cancho V G and Cordeiro G M. 2010. *Computational Statistics and Data Analysis*, 54 1017-1035
[9] Ortega E M M, Cordeiro G M and Kattan M W 2013 *Stat Papers*. 54 113-132
[10] Mahmoud M R, Morshera N A, and Ahmad M A M 2016 *American Journal of Applied Mathematics and Statistics*. 4(3) 74-86
[11] Mayo Clinic 2000 Mayo Clinic Primary Biliary Cirrhosis Data [cited 2019 Jan 30]. Available from: https://vincentarelbundock.github.io/Rdatasets/datasets.html
[12] Mohammed A, Naugler C and Far H 2015 *Emerging Trends in Computer Science and Applied Computing*. 577-602