IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/rjes

Bayesian method estimation for Exponential and Weibull Survival regression models

Wadhah S. Ibrahim [1], Ahmed Salam Mezher [2]

[1,2] Department of Statistics, College of Administration and Economics, Al_Mustansiriyah University Baghdad, Iraq.

ARTICLE INFO

Article history:
Received 13 February 2024
Revised 13 February 2024,
Accepted 02 February 2024,
Available online 02 February 2024

Keywords:
Cox regression model,
Exponential survival regression model,
Weibull survival regression model,
Bayesian method,
Bayesian Information Criterion (BIC).

ABSTRACT

Parametric regression models are one of the most important regression models used in the medical field. They are the tool through which the response variable is modeled when the values of that variable are survival times with a known probability distribution such as the exponential distribution and the Weibull distribution, I propose in this article to estimate the parameters of parametric survival regression models using the Bayesian method, these models are: the exponential survival regression model, the Weibull survival regression model, The simulation was used to generate data that follows the parametric survival regression models depending on various factors such as sample size, and three different parameter models, The simulation results showed that the exponential survival regression model outperformed the Weibull survival regression model in order to obtain the lowest value according to the Bayesian Information Criterion (BIC), The larger the sample size, the more accurate and reliable the analysis estimators performed by the Bayes Information Criterion.

1. Introduction

Survival analysis refers to the analysis of elapsed timer. The response variable is the relative time between the beginning of the event and its end, whereas the end time of the event is either the time of occurrence of the event under study, such as death, failure, or the end of the individual follow-up, and the study of the variable of elapsed times It has two characteristics. The first is that traditional statistical methods such as the t-test, data analysis, and regression models are not appropriate for analysis because all the elapsed times are positively skewed, meaning that most of the data is concentrated on the right side of the distribution, Statistical methods require data that is normally distributed, and this is not available in survival data. The second is that these data are incomplete (Censored) when the end event occurs and they are of several different types, and when attention is focused on studying the variable of death, age, the testing period, as well as the variable Survival time is calculated after the patient has been diagnosed with a specific disease such as cancer and under normal conditions for treatment, Among the statistical methods that are concerned with studying the relationship between the response variable represented by survival time and a group of explanatory variables is a multiple regression model. Unfortunately, due to the specific nature of survival data, the regression model is not appropriate, since survival data contain incomplete data (Censored data) of many types, in addition to the fact that the distribution of survival times is often highly skewed, these two problems are acceptable and therefore multiple regression models cannot be used. Several methods have been proposed. For study, the most famous and most appropriate method for constructing survival regression models is the Cox regression model, which was developed by the English scientist Cox dived in

* Corresponding author.
E-**mail** address: dr_wadhah_stat@uomustansiriyah.edu.iq

1972, through which the survival time of the injured person and the factors affecting survival time are studied. It was used in analyzing medical experiments through the effect of the number of Variables on length of stay [5,7].

There are many survival regression models that deal with this type of study, including parametric and semi-parametric. These methods are related to analyze survival data, which differ according to the nature of the phenomenon studied. Given the importance of these models in scientific life, many researchers have conducted studies from many different points of view, and researchers are still interested in studying this. Models exist so far. Parametric survival models are among the most widely used models, and it is assumed that the survival time follows a known distribution, and the commonly used parametric survival distributions are (Exponential, Weibull) and others [6].

2. Cox Regression Model

The following formulas represent the hazard rate function, the probability density function, and the survival function for the Cox model:

$$h_i(t) = h_0(t)exp(\beta' X) \quad ...(1)$$

$$S_i(t) = \left(S_0(t)\right)^{exp(\beta' X)} \quad ...(2)$$

Where: $h_0(t)$ and $S_0(t)$ is The Hazard Function, and the survival function for the distribution under study [10,11].

3. Exponential Survival Regression Model

It is a continuous probability distribution that is widely used because it is one of the distributions with simple application and is described as a model of a constant failure rate over time, as it is used to estimate the time periods for the occurrence of events represented by the failure and success of the units or samples included in the test [1, 13].

The probability density function (p.d.f) for this model is [2]:

$$f_i(t) = \lambda exp(-\lambda t) \;\; t > 0 \,, \;\; \lambda > 0 \quad ...(3)$$

Since λ: the Scale Parameter.

Its survival function $S_i(t)$ is [9]:

$$S_i(t) = e^{-\lambda t} \,, \quad h_i(t) = \lambda \,, \; H_i(t) = \lambda t \,, \; F_i(t) = 1 - e^{-\lambda t} \quad ...(4)$$

$h_i(t)$: The Hazard Function, $H_i(t)$: The Hazard Function Cumulative, $F_i(t)$: The Cumulative Density Function[8].

Therefore, the exponential regression model will be defined as follows according to equation (1) [11].

$$h_i(t) = \lambda exp(\beta' X) \quad ...(5)$$

The survival function can be written based on the relative risk model if it is assumed that the survival time follows an exponential distribution in the following form according to the equation (1).

$$S_i(t) = \left(exp(-\lambda t)\right)^{exp(\beta' X)} \quad ...(6)$$

Generating data for an exponential survival regression model using the inverse transformation method as follows [3]:

$$t_i = \frac{-\ln(U)}{\lambda exp(\beta' X)} \quad ...(7)$$

4. Weibull Survival Regression Model

It is one of the most important models because it is characterized by the risk function varying between increasing, decreasing, and constant due to the different values of the model parameters [14].

The probability density function (p.d.f) of the two-parameter Weibull distribution is [12]:

$$f_i(t) = a\lambda(\lambda t)^{a-1}\exp\left[-(\lambda t)^a\right] \quad ...(8)$$

Since λ: the Scale Parameter , a: the Shape parameter.

$$S_i(t) = \exp[-(\lambda t)^a] \,, h_i(t) = a\lambda(\lambda t)^{a-1}$$

$$H_i(t) = (\lambda t)^a \,, F_i(t) = 1 - e^{-(\lambda t)^a} \quad ...(9)$$

Since The formulas are already defined.

Accordingly, the Weibull regression model will be defined according to equation (1) [4]:

$$h_i(t) = \alpha\lambda(\lambda t)^{\alpha-1}exp(\beta' X) \quad ...(10)$$

The survival function can be written based on the relative risk model in general according to equation (1):

$$S_i(t) = \left(exp(-(\lambda t)^{\alpha})\right)^{exp(\beta' X)} \quad ...(11)$$

Data for the Weibull survival regression model can be generated using the back transformation method as follows p3]:

$$t_i = \frac{1}{\lambda}\left[\frac{-\ln(U)}{exp(\beta' X)}\right]^{1/\alpha} \quad \dots(12)$$

5. Estimating the parameters of the exponential survival regression model using the Bayesian method

We assume a probability distribution (The Prior distribution) for the parameter (λ_i) that has the following formulas [7,16]:

$$\lambda_i \sim N(\mu_\lambda, \sigma_\lambda^2) \quad i = 1, \dots, n \quad \dots(13)$$

Since the parameter (σ_λ^2) is a variance of the parameter (λ_i), we can say, according to Jeffery's rule, that it has an inverse chi-square distribution, that is [10]:

$$P\left(\frac{\sigma_\lambda^2}{y_\lambda}\right) \propto \left(\sigma_\lambda^2\right)^{-\left(\frac{y_\lambda}{2}+1\right)} \exp\left(\frac{-y_\lambda}{2\sigma_\lambda^2}\right) \quad \dots(14)$$

The joint initial probability density function is as follows:

$$\int_{-\infty}^{\infty} exp - \left[\frac{n(\bar{\lambda}-\mu_\lambda)^2}{2\sigma_\lambda^2}\right] d\mu_\lambda \propto \left(\sigma_\lambda^2\right)^{1/2} \quad \dots(15)$$

The joint posterior distribution of the parameter (λ_i) is determined as follows [16]:

$$\int_0^{\infty} \left(\sigma_\lambda^2\right)^{-(n+y_\lambda+1)/2} exp\left[\frac{-\left[y_\lambda+\sum_i(\lambda_i-\bar{\lambda})^2\right]}{2\sigma_\lambda^2}\right] d\sigma_\lambda^2 \propto \left[y_\lambda + \sum_i(\lambda_i-\bar{\lambda})^2\right]^{-(n+y_\lambda-1)/2} \quad \dots(16)$$

Using the equation, the maximum likelihood function for the exponential survival regression model:

$$P(\lambda_i/t) = \lambda^n exp\left(\sum_{i=1}^{n}\beta' X\right)\prod_{i=1}^{n}\left(exp(-\lambda t_i)\right)^{exp(\beta' X)} * \left[y_\lambda + \sum_i(\lambda_i-\bar{\lambda})^2\right]^{-(n+y_\lambda-1)/2} \quad \dots(17)$$

Integration of the joint posterior distribution above is inappropriate despite making it a normal distribution. By taking the natural logarithm and then finding the derivative and setting it equal to zero, the formula will be as follows:

$$f(\lambda_i) = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} exp(\beta' X)(t_i) + \left(\hat{\lambda_i} - \hat{\lambda}\right)/\vartheta_\lambda \quad \dots(18)$$

$$z(\beta) = \sum_{i=1}^{n} X - \lambda \sum_{i=1}^{n} exp(\beta' X) X(t_i) \quad \dots(19)$$

The first derivative of functions (18) and (19) is as follows:

$$\acute{f}(\lambda_i) = -\frac{n}{\hat{\lambda}^2} + \left[\vartheta_\lambda\left(1-\frac{1}{n}\right) - 2\left(\hat{\lambda_i} - \hat{\lambda}\right)^2/(n+y_\lambda-1)\right]/(\vartheta_\lambda)^2 \quad \dots(20)$$

$$\acute{z}(\beta) = -\lambda \sum_{i=1}^{n} exp(\beta' X) X^2(t_i) \quad \dots(21)$$

We use the Newton-Raphson method to find the estimators as follows [7]:

$$\hat{\lambda}_i^{t+1} = \hat{\lambda}_i^t - \frac{f(\hat{\lambda}_i^t)}{\acute{f}(\hat{\lambda}_i^t)} \;, \hat{\beta}_i^{t+1} = \hat{\beta}_i^t - \frac{z(\hat{\beta}_i^t)}{\acute{z}(\hat{\beta}_i^t)} \quad \dots(22)$$

6. Estimating the parameters of the Weibull survival regression model using the Bayesian method

We assume a probability distribution (The Prior distribution) for the parameters (a_i, λ_j) that has the following formulas [7,16]:

$$a_i \sim N(\mu_a, \sigma_a^2) \quad i = 1, \dots, n$$
$$\lambda_j \sim N(\mu_\lambda, \sigma_\lambda^2) \quad j = 1, \dots, m \quad \dots(23)$$

Since the parameters $(\sigma_a^2, \sigma_\lambda^2)$ are the variance of each of the parameters (a_i, λ_j), we can say, according to Jeffery's rule, that it has an inverse chi-square distribution, that is [10]:

$$P\left(\frac{\sigma_a^2}{y_a}\right) \propto \left(\sigma_\lambda^2\right)^{-\left(\frac{y_a}{2}+1\right)} exp\left(\frac{-y_a}{2\sigma_a^2}\right) \quad \dots(24)$$

$$P\left(\frac{\sigma_\lambda^2}{y_\lambda}\right) \propto \left(\sigma_\lambda^2\right)^{-\left(\frac{y_\lambda}{2}+1\right)} exp\left(\frac{-y_\lambda}{2\sigma_\lambda^2}\right) \quad \dots(25)$$

The joint initial probability density function is as follows:

$$\int_{-\infty}^{\infty} exp - \left[\frac{n(\bar{a}-\mu_a)^2}{2\sigma_a^2}\right] d\mu_a \propto (\sigma_a^2)^{1/2} \dots(26)$$

$$\int_{-\infty}^{\infty} exp - \left[\frac{m(\bar{\lambda}-\mu_\lambda)^2}{2\sigma_\lambda^2}\right] d\mu_\lambda \propto \left(2\sigma_\lambda^2\right)^{1/2} \dots(27)$$

The joint posterior distribution is determined for the parameters (a_i, λ_j), as follows [16]:

$$\int_0^{\infty} (\sigma_a^2)^{-(n+y_a+1)/2} exp\left[\frac{-[y_a + \sum_i(a_i-\bar{a})^2]}{2\sigma_a^2}\right] d\sigma_a^2 \propto \left[y_a + \sum_i (a_i-\bar{a})^2\right]^{-(n+y_a-1)/2} \quad \dots(28)$$

$$\int_0^{\infty} (\sigma_\lambda^2)^{-(m+y_\lambda+1)/2} exp\left[\frac{-\left[y_\lambda + \sum_j(\lambda_j-\bar{\lambda})^2\right]}{2\sigma_\lambda^2}\right] d\sigma_\lambda^2$$

$$\propto \left[y_\lambda + \sum_i \left(\lambda_j - \bar{\lambda}\right)^2 \right]^{-(m+y_\lambda-1)/2} \quad \dots (29)$$

Using the equation for the maximum likelihood function of the Weibull survival regression model $P(a_i, \lambda_j / t)$:

$$= \alpha^n \lambda^{\alpha n} exp\left(\sum_{i=1}^{n} \beta' X\right) \prod_{i=1}^{n} (t_i)^{\alpha-1} \left(exp(-(\lambda t_i)^\alpha)\right)^{exp(\beta' X)}$$

$$[y_a + \sum_i (a_i - \bar{a})^2]^{-(n+y_a-1)/2} \left[y_\lambda + \sum_i \left(\lambda_j - \bar{\lambda}\right)^2 \right]^{-(m+y_\lambda-1)/2} \quad \dots (30)$$

Integration of the joint posterior distribution above is inappropriate despite making it a normal distribution. By taking the natural logarithm and then finding the derivative and setting it equal to zero, the equations will be as follows [7]:

$$f(a_i) = \frac{n}{\hat{\alpha}} + \frac{n}{\hat{\lambda}} + \frac{1}{\sum_{i=1}^{n}(t_i)} + \frac{\sum_{i=1}^{n} exp(\beta' X)((\lambda t_i)^\alpha)}{\hat{\lambda} t_i} - \frac{(\hat{a}_i - \hat{\bar{a}})}{\vartheta_\alpha} \quad \dots (31)$$

$$h(\lambda_j) = \frac{\alpha n}{\hat{\lambda}} + \sum_{i=1}^{n} exp(\beta' X)\, \hat{\alpha} t^a \hat{\lambda}^{a-1} - \left(\hat{\lambda}_j - \hat{\bar{\lambda}}\right)/\vartheta_\lambda \ , \ z(\beta) = \sum_{i=1}^{n} X + \sum_{i=1}^{n} exp(\beta' X) X\, ((\lambda t_i)^\alpha) \quad \dots (32)$$

The first derivative of functions (31) and (32) is as follows:

$$\acute{f}(a_i) = -\frac{n}{\hat{\alpha}^2} + \frac{\sum_{i=1}^{n} exp(\beta' X)((\hat{\lambda} t_i)^{\hat{\alpha}})}{\hat{\lambda} t_i^{\,2}} + \left[\vartheta_\alpha \left(1 - \frac{1}{n}\right) - 2(\hat{a}_i - \hat{\bar{a}})^2/(n + y_a - 1)\right]/(\vartheta_\alpha)^2 \quad \dots (33)$$

$$\acute{h}(\lambda_j) = -\frac{an}{\hat{\lambda}^2} + \sum_{i=1}^{n} exp(\beta' X)\, \hat{\alpha} t^a (\hat{\alpha} - 1) \hat{\lambda}^{a-2} + \left[\vartheta_\lambda \left(1 - \frac{1}{m}\right) - 2\left(\hat{\lambda}_j - \hat{\bar{\lambda}}\right)^2/(m + y_\lambda - 1)\right]/(\vartheta_\lambda)^2 \quad \dots (34)$$

$$\acute{z}(\beta) = \sum_{i=1}^{n} exp(\beta' X) X^2\, ((\lambda t_i)^\alpha) \dots (35)$$

We use the Newton-Raphson method to find the estimators as in equation (22).

7. Bayesian Information Criteria

It is the criterion for choosing the best model from among several statistical models, as this criterion selects the simplest statistical models. Its mathematical formula is [8,9]:

$$BIC = -2 \log L + (a) \log(n) \quad \dots (36)$$

8. Simulation study

The statistical programming language R 4.3.1 was used to write the simulation program. The written program includes four basic stages for estimating parametric Cox models, as follows [10.13]:

1. Determining initial values for parameters

Table (1-1): Initial values for regression parameters

β	Model I	Model II	Model III
β1	0.00138	0.00153	0.00168
β2	-0.28910	-0.26282	-0.23654
β3	0.01715	0.01906	0.02097
β4	-0.11220	-0.10200	-0.09180
β5	0.38332	0.42591	0.46850
β6	-0.08032	-0.07302	-0.06572
β7	0.01155	0.01283	0.01411
β8	-0.21980	-0.19982	-0.17984
β9	-0.06155	-0.05595	-0.05036

Different values have been assumed for the parameters of each of the distributions used, as follows:

Table (1-2): Initial values for parameters

Distribution	**λ**	**α**
Exponential	0.1	-
	0.5	-
	1	-
Weibull	0.1	0.5
	0.5	1.5
	1	2.5

Three different sample sizes were chosen: (50, 100, 200). As for the repetition of each of these experiments, it was repeated a thousand times.

2. Data generation: Each of the explanatory variables is generated from a uniform distribution U (0, 1) using the runif function in the R statistics package, and survival times for each of the used distributions are generated using the specified formulas. In equations (7) and (12).

3. Estimates: At this stage, the estimation process is performed for the parameters of the distributions used using the Bayesian method, as well as the estimation of the regression parameters.
4. Comparison between models: For the purpose of comparing the behaviour of the models used, the Bayesian Information Criteria (BIC) was used, as defined in equation (36).

Simulation results: The results were as follows:

Table (1-3): the BIC for the Exponential survival regression model.

Parameters	n	BIC		
		Model I	Model II	Model III
$\lambda = 0.1$	50	41.678413	40.191504	38.695326
	100	48.668682	47.203604	45.721145
	200	55.649272	54.171106	52.696487
$\lambda = 0.5$	50	38.447122	36.945708	35.444833
	100	45.451205	43.974230	42.477675
	200	52.446265	50.961359	49.475504
$\lambda = 1$	50	37.038632	35.570131	34.100710
	100	51.047284	49.557779	48.096867
	200	55.433956	54.904279	52.465628

Table (1-4): the BIC for the Weibull survival regression model.

Parameters	n	BIC		
		Model I	Model II	Model III
$\lambda = 0.1$ $\alpha = 0.5$	٥٠	41.831488	38.887800	35.934171
	١٠٠	49.517418	46.583285	43.667233
	٢٠٠	57.260163	54.279073	51.322470
$\lambda = 0.5$ $\alpha = 1.5$	50	84.593282	70.320387	48.912659
	100	89.619510	75.108044	64.128331
	200	95.496573	87.637709	80.075795
$\lambda = 1$ $\alpha = 2.5$	50	63.687342	43.389374	37.475598
	100	70.252035	47.635428	41.178692
	200	77.740678	69.933806	67.253902

Table (١-٥): the BIC for comparison between the models used.

Distr.	Parameters	n		
		50	100	200
Exponential	$\lambda = 0.1$	38.69532	45.72114	52.69648
	$\lambda = 0.5$	35.44483	42.47767	49.47550
	$\lambda = 1$	34.10071	48.09686	52.46562
Weibull	$\lambda = 0.1$ $\alpha = 0.5$	35.93417	43.66723	51.32247
	$\lambda = 0.5$ $\alpha = 1.5$	48.91265	64.12833	80.07579
	$\lambda = 1$ $\alpha = 2.5$	37.47559	41.178692	67.253902

9. Conclusions

The most important conclusions were:

1. The best model used is the exponential model because it has the lowest value according to the (BIC), in 9 cases out of a total of 9 cases studied.
2. The estimators of the (BIC) increase with increasing sample size at different values of the default parameters of the distribution. Simply put, the larger the sample size, the more accurate and reliable the estimators of the analysis performed by the Bayesian Information Criterion.
3. Bayesian BIC estimators decrease in value between the three models used and for different sample sizes.

References

[1] Al-Suhail, Aseel Mahmoud Shaker (2016) "Estimating the reliability of systems using nonparametric and semi-parametric Bayesian estimators with a practical application" *College of Administration and Economics* - University of Baghdad.

[2] Al-Tanja, Main (2014) "Finding the lowest possible risk in the Cox regression model," PhD thesis, University of Aleppo, Syria.

[3] Bender, R., Augustin, T., & Blettner, M. (2005). "Generating survival times to simulate Cox proportional hazards models". *Statistics in medicine*, 24(11), 1713-1723.

[4] Cleves, M. (2008). “An introduction to survival analysis using Stata”. Stata press.

[5] Collett, D. (2023). “Modelling survival data in medical research”.4th Edition CRC press.

[6] Gui, J., & Li, H. J. B. (2005). “Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data”. 21(13), 3001-3008.

[7] Ibrahim, Wadhah S., “Methods of Estimating the RASCH Model for Multiple Categorical Data Measurements with practical application”, Ph.D. dissertation, Dept. of Stat., College of Administration and Economics, Baghdad University, Iraq, 2016.

[8] Ibrahim, Wadhah S. , Aliwi, Ali K.,(2023) "using the method of maximum likelihood estimation to estimate the survival function of the new expanded transformer Weibull distribution", *Journal of Statistical Science* 19, 23-33.

[9] Ibrahim, Wadhah S. , Khaleel, Basheer j.,(2023), " Comparison of Some Classical Methods for Estimating the Survival Function of the Two-Parameter Lindley Distribution", *Journal of Statistical Sciences* 17, 70-81.

[10] Ibrahim, Wadhah S., Mhadi, Dijla I. (2016), " A Comparison of some methods for estimating Rasch model parameters" *journal of the college of basic education* 22 (93), 245-260.

[11] Karim, M. R., & Islam, M. A. (2019). “Reliability and survival analysis”. Springer Singapore.`

[12] Lee, E., Wang, J. (2003)." Statistical methods for survival data analysis". Third edition, Wiley.

[13] Liu, Xian., (2012). "Survival analysis: models and applications", John Wiley & Sons Ltd, The Atrium, Southern Gate, Chi Chester, West Sussex, PO19 8SQ, United Kingdom.

[14] Makhoul Mtanios and Ghanem Adnan 2011, “The Effectiveness of Using the Weibull Probability Distribution for Prediction”, Damascus University Journal of Economic and Legal Sciences, Volume 27, Issue 4, pp. 138-199.

[15] Muse, A. H., Ngesa, O., Mwalili, S., Alshanbari, H. M., & El-Bagoury, A. A. H. (2022). “A flexible Bayesian parametric proportional hazard model: Simulation and applications to right-censored healthcare data”. Journal of Healthcare Engineering.

[16] Swami Nathan, H., & Gifford, J. A. (1982). “Bayesian estimation in the Rash model”. *Journal of Educational Statistics*, 7(3), 175-191.