مجلة كلية التربية الاساسية
كلية التربية الاساسية – الجامعة المستنصرية
Journal of the College of Basic Education Vol.30 (NO. 124) 2024, pp. 1-10

Using the Bayesian method to estimate and comparison the regression of the exponential and gamma survival (Simulation)

Ahmed Salam Mezher **Wadhah S. Ibrahim**
Al-Mustansiriya University - Economics and Administration
Department of Statistics
drwadhah_stat@uomustansiriyah.edu.iq ahmed.salam@uomustansiriyah.edu.iq

Abstract:

Parametric regression models It is one of the oldest and most common regression models and can be defined as one of the statistical methods that is used to describe and estimate the relationship between a dependent random variable and explanatory random variables. These models have types, including quantitative models, of their linear and non-linear types, and qualitative models that The dependent variable is a binary response variable. These models are characterized by the fact that all cases that are studied are assumed to be normally distributed and measurable, and that the regression function determines the parameters that cannot be changed except by changing the number of variables included in the study, relying on probability distributions, which are the exponential distribution and the Gamma, we will obtain parametric regression models, which are the exponential regression model, and the gamma regression model, relying on the Cox regression model to be used in forming these models, which are models that require the presence of information and hypotheses about the distribution of the population used in the test that is appropriate for the units or data, as well as the variables that enter. The test or experiment on which the survival function is based, in this thesis, it was proposed to estimate the parameters of these models using the Bayesian method. The simulation method was used to generate data that follows parametric survival regression models according to various factors. The simulation results showed that the third model of the exponential probability distribution of the survival function shows the lowest value according to the MSE criterion. This indicates that Increasing the sample size typically reduces the standard mean error (MSE), which is used as an indicator of the differences between expected and actual values in statistical estimates. Simply put, the larger the sample size, the more accurate the statistical predictions are and the lower the standard deviation from the actual values, indicating that larger samples often contribute to improving the accuracy of statistical estimates.

Keywords: Cox regression model, exponential survival regression model, Gamma survival regression model, Bayesian method, Mean Squared Error criterion (mse).

1. Introduction

Survival analysis is considered one of the most important modern analysis methods, and it is just another name for analyzing the time until an event occurs (event probability). The term survival analysis is mostly used in biomedical sciences, where the interest is in monitoring the time to death for either patients or laboratory animals. Time analysis has also been used Although the event is widely used in social sciences where interest is in analyzing the time of events such as job changes, marriage, birth of children, etc., engineering sciences also contributed to the development of survival analysis called “reliability analysis”, or “failure time analysis”. In this field, because the main focus is on modeling the time it takes for machines or electronic components to fail, developments from these diverse fields have been mostly combined in the field of “survival analysis.” (Fox & Weisberg, 2002).

The study of survival analysis has taken on a large role among researchers, especially in recent years, because it plays a major role in many areas of life, such as medicine and biological fields, and because of its great importance in studying the survival of living organisms after a specific period of time (t Survival analysis plays an important role in identifying risk factors in the biomedical and industrial fields. Given the great interest that the subject of survival analysis has received from researchers, these studies have become courses taught at various levels of study, and thus survival analysis has become an independent science that plays a fundamental role in analyzing the event. Estimation, prediction, and optimization.. (Karim, Atheer Abdel Zahra 2018).

2. Aim of research

The research aims to clarify the survival regression model using different distributions (exponential, gamma) and employ these distributions in the Cox regression model to form survival regression models, use the Bayesian method to estimate the parameters of these models, and compare between those models using the simulation method using MSE.

3. Cox Regression Model

is considered one of the most important and most common models in survival analysis models that deal with time in the analysis, as this method has many advantages, the most important of which is that it is one of the methods characterized by the accuracy of its results, as well as the ease of dealing with disappearance data that appears when time is taken into account. It is one of the proportional hazards models. (Fox 2002).

$$\frac{h_i(t)}{h_0(t)} = exp(\eta_i), \quad \eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \quad ...(1)$$

Where p represents the number of explanatory variables, x_{ji} represents the value of the ith observation from the jth-sequenced explanatory variable, β_j is the regression parameter for the explanatory variable, and $h_0(t)$ represents the hazard function of the probability distribution. It is noted that the Cox model does not contain β_0 , and this is due to This is because $h_0(t)$ replaces it, The following formulas represent the hazard rate function, and the survival function for the Cox model: (Karim & Islam, 2019: 120-121) (Muse et al., 2022: 3-4).

$$\mathbf{h_i(t) = h_0(t)exp(\eta_i)\,, S_i(t) = (S_0(t))^{exp(\eta_i)}} \quad ...(2)$$

4. Exponential survival regression model

One of the advantages of this distribution is that its hazard rate is a fixed amount, and this indicates that the causes of failure are not due to obsolescence, but rather due to random accidents. The parametric exponential regression model is one of the easiest parametric survival models, because it assumes that the survival time follows an exponential distribution. (Abdel Moneim, Tharwat 2012)

The probability density function (p.d.f) for this model is: (Al-Tanja, Main 2014).

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

Its survival function $S_i(t)$ is: (Ibrahim, Wadhah & Khaleel, Basheer,2023).

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

$h_i(t)$: The Hazard Function, $H_i(t)$: The Hazard Function Cumulative, $F_i(t)$: The Cumulative Density Function. (Ibrahim, Wadhah & Aliwi, Ali 2023).

Therefore, the exponential regression model will be defined as follows according to equation (2).

$$h_i(t) = \lambda_e \mathbf{exp(\eta_i)} \quad ...(5)$$

The survival function can be written based on the relative hazard model if it is assumed that the survival time follows an exponential distribution in the following form according to the equation (2).

$$S_i(t) = \left(exp(-\lambda_e t_i)\right)^{\exp(\eta_i)} \quad \dots(6)$$

Generating data for an exponential survival regression model Using the inverse transformation method as follows: (Bender et al., 2005).

$$t_{i_e} = \frac{-\ln(U)}{\lambda_e \exp(\eta_i)} \quad \dots(7)$$

5. Gamma survival regression model

The gamma distribution is considered a continuous probability distribution, as the name was derived from the famous gamma function. This distribution is used to measure lead times and waiting periods. (Haerting et al., 2007)

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

$$f_i(t) = \frac{\lambda(\lambda t)^{a-1}\, e^{-\lambda t}}{\Gamma(a)} \quad \dots(8)$$

$$\gamma(\alpha, \lambda t) = \frac{1}{\Gamma(a)} \int_0^{\lambda t} u^{a-1}\ e^{-u}\, du \quad \dots(9)$$

Since λ: the Scale Parameter , α: the Shape parameter, $\gamma(\alpha, \lambda t_i)$: incomplete gamma function.

$$S_i(t) = 1 - \gamma(\alpha, \lambda t), h_i(t) = \frac{\lambda(\lambda t)^{\alpha-1}\, e^{-\lambda t}}{\Gamma\alpha - \gamma(\alpha, \lambda t)}, F_i(t) = \gamma(\alpha, \lambda t), H_i(t) = -\ln S_i(t) = -\ln\left(1 - \gamma(\alpha, \lambda t)\right) \quad \dots(10)$$

Since The formulas are already defined. (Haerting et al., 2007)

Accordingly, the Weibull regression model will be defined according to equation (2):

$$h_i(t) = \frac{\lambda(\lambda t)^{\alpha-1}}{\Gamma\alpha - \gamma(\alpha, \lambda t)} exp(\eta_i - \lambda t) \quad \dots(11)$$

The survival function can be written based on the relative hazard model in general according to equation (2):

$$S_i(t) = \left(\frac{\Gamma\alpha - \gamma(\alpha, \lambda t)}{\Gamma\alpha}\right)^{\exp(\eta_i)} \quad \dots(12)$$

Data for the Weibull survival regression model can be generated using the back transformation method as follows: (Bender et al., 2005).

$$t_{i_G} = \frac{1}{\lambda} \gamma^{-1}\left(\alpha, \Gamma\alpha - exp\left[\frac{ln(U)}{\exp(\eta_i)} + ln(\Gamma\alpha)\right]\right) \quad \dots(13)$$

6. 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_e}) that has the following formulas: (Swami Nathan, Gifford, J. 1982)

$$\lambda_{i_e} \sim N(\mu_\lambda, \sigma_\lambda^2) \quad i = 1, \dots, n \qquad \dots(14)$$

Since the parameter (σ_λ^2) is a variance of the parameter (λ_{i_e}), we can say, according to Jeffery's rule, that it has an inverse chi-square distribution, that is: (Ibrahim, Wadhah & Mhadi, Dijla 2016)

$$P\left(\frac{\sigma_\lambda^2}{y_\lambda}\right) \propto (\sigma_\lambda^2)^{-\left(\frac{y_\lambda}{2}+1\right)} \exp\left(\frac{-y_\lambda}{2\sigma_\lambda^2}\right) \qquad \dots(15)$$

The joint posterior distribution of the parameter (λ_{i_e}) is determined as follows: (Swami Nathan & Gifford, J. 1982)

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

$$\left[y_\lambda + \sum_i (\lambda_i - \bar{\lambda})^2\right]^{-(n+y_\lambda-1)/2} \qquad \dots(16)$$

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

$$P(\lambda_i/t) = \lambda_e{}^n exp\left(\sum_{i=1}^{n} \beta' X\right) \prod_{i=1}^{n} (exp(-\lambda_e t_i))^{exp(\beta' X)}$$

$$* \left[y_\lambda + \sum_i (\lambda_i - \bar{\lambda})^2\right]^{-(n+y_\lambda-1)/2} \qquad \dots(17)$$

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: (Swami Nathan & Gifford, J. 1982)

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

$$w(\beta) = \sum_{i=1}^{n} X - \lambda_e \sum_{i=1}^{n} exp(\beta' X) X(t_i) \qquad \dots(19)$$

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

$$\acute{f}(\lambda_i) =$$

$$-\frac{n}{\widehat{\lambda_e}^2} + \left[\vartheta_\lambda \left(1 - \frac{1}{n}\right) - 2\left(\widehat{\lambda_\iota} - \hat{\bar{\lambda}}\right)^2 / (n + y_\lambda - 1)\right] / (\vartheta_{\lambda_e})^2 \qquad \dots(20)$$

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

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

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

7. Estimating the parameters of the Gamma survival regression model using the Bayesian method

Applying the same previous steps of the Bayesian method above, we will start from the maximum likelihood function for the gamma survival regression model as follows: (Swami Nathan, Gifford, J. 1982)

$$P(a_i, \lambda_j / t) = \frac{\lambda^{n\alpha}}{(\Gamma\alpha)^n} exp\left(\sum_{i=1}^{n} \beta' X\right) \prod_{i=1}^{n} (t_i)^{\alpha-1} exp(-\lambda t_i) \left(\frac{\Gamma\alpha - \gamma(\alpha, \lambda t)}{\Gamma\alpha}\right)^{exp(\beta' X)-1} \left[y_a + \sum_i (a_i - \bar{a})^2\right]^{-(n+y_a-1)/2} * \left[y_\lambda + \sum_i (\lambda_j - \bar{\lambda})^2\right]^{-(m+y_\lambda-1)/2} \quad \dots(23)$$

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: (Swami Nathan, Gifford, J. 1982)

$$f(a_i) = \frac{n}{\hat{\lambda}} - \frac{n - 1 + \sum_{i=1}^{n}(exp(\beta' X) - 1)}{\Gamma\hat{a}} + \frac{1}{\sum_{i=1}^{n}(t_i)} + (\hat{a}_i - \hat{\bar{a}})/\vartheta_\alpha \quad \dots(24)$$

$$h(\lambda_j) = \frac{n\hat{a}}{\hat{\lambda}} - \sum_{i=1}^{n} t_i - \left(\hat{\lambda}_j - \hat{\bar{\lambda}}\right)/\vartheta_\lambda \quad , \quad z(\beta) = \sum_{i=1}^{n} X + \frac{(exp(\beta' X) - 1)X}{\Gamma\hat{a}} \quad \dots(25)$$

The first derivative of functions (24) and (25) is as follows:

$$\acute{f}(a_i) = \frac{n - 1 + \sum_{i=1}^{n}(exp(\beta' X) - 1)}{\Gamma\hat{a}^2} + \left[\vartheta_\alpha\left(1 - \frac{1}{n}\right) - 2(\hat{a}_i - \hat{\bar{a}})^2/(n + y_a - 1)\right]/(\vartheta_\alpha)^2 \quad \dots(26)$$

$$\acute{h}(\lambda_j) = -\frac{n\hat{a}}{\hat{\lambda}^2} + \left[\vartheta_\lambda\left(1-\frac{1}{m}\right) - 2\left(\hat{\lambda}_j - \hat{\hat{\lambda}}\right)^2 / (m + y_\lambda - 1)\right] / (\vartheta_\lambda)^2 , \quad \acute{z}(\beta) = \frac{(exp(\beta' X) - 1)X^2}{\Gamma\hat{\alpha}} \quad \ldots(27)$$

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

8. Mean Squared Error criterion (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{S}(t_i) - S(t_i)\right)^2 \quad \ldots(28)$$ (Box, G. E. & Tiao 1979)

9. 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:

1. Determining the initial values of the parameters: The initial values of the regression B parameters were determined by three different models (Model I, Model II, and Model III), and different values were assumed for the parameters of each of the distributions used (λ=1, 0.5, 0.1) and (α=0.5, 1.5, 2.5). Three different 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 (13).
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 behavior of the models used, the Mean Squared Error Criteria was used, as defined in Equation. (28).

Simulation results: The results were as follows:

Table (1-3): MSE values for the survival function for the distributions used when λ = 0.5, α = 0.5

Sample Size	Values	Model I	Model II	Model III
n=50	Exponential	0.01509015	0.0150131	0.0150995
	Gamma	0.0154928	0.0157474	0.0156157
n=100	Exponential	0.00709523	0.0071303	0.0073186
	Gamma	0.00716212	0.0072409	0.0072676
n=200	Exponential	0.00347957	0.0034345	0.0033613
	Gamma	0.00347115	0.0033537	0.0034579

Table (1-4): MSE values for the survival function for the distributions used when λ = 0.5, α = 1.5

Sample Size	Values	Model I	Model II	Model III
n=50	Exponential	0.01530092	0.0147286	0.0147611
	Gamma	0.01492227	0.0151048	0.0146692
n=100	Exponential	0.00691752	0.0068569	0.0070901
	Gamma	0.00691584	0.0069803	0.0069655
n=200	Exponential	0.00335175	0.0033532	0.0033371
	Gamma	0.00339671	0.0033363	0.0033982

Table (1-5): MSE values for the survival function for the distributions used when λ = 0.5, α = 2.5

Sample Size	Values	Model I	Model II	Model III
n=50	Exponential	0.015191	0.0148846	0.0149239
	Gamma	0.01470421	0.0145254	0.014632
n=100	Exponential	0.00700175	0.0070232	0.0070437
	Gamma	0.00689876	0.0068121	0.006993
n=200	Exponential	0.00340076	0.0033376	0.0032136
	Gamma	0.00326764	0.0032143	0.0032599

Table (1-6): MSE values for comparison between the survival function of the distributions used

Parameters	Distribution	n		
		50	100	200
$\lambda = 0.5, \alpha = 0.5$	Exponential	0.0150131	0.00709523	0.0033613
	Gamma	0.0154928	0.00716212	0.0033537
$\lambda = 0.5, \alpha = 1.5$	Exponential	0.0147286	0.0068569	0.0033371
	Gamma	0.0146692	0.00691584	0.0033363
$\lambda = 0.5, \alpha = 2.5$	Exponential	0.0148846	0.00700175	0.0032136
	Gamma	0.0145254	0.0068121	0.0032143

10. Conclusions

The most important conclusions were:

1. We conclude from the results of Table No. (1-3), (1-4) that the survival function of the second model of the gamma distribution is preferable at a sample size of 200 in order to obtain the lowest value according to the MSE criterion, as shown in the table above.
2. We conclude from the results of Table No. (1-5) that the survival function of the third model of the exponential distribution is preferable at a sample size of 200 in order to obtain the lowest value according to the MSE criterion, as shown in the table above.
3. From Table No. (1-6), we can conclude that the third model of the exponential probability distribution of the survival function shows the lowest value in the previous table according to the MSE criterion. This indicates that increasing the sample size usually reduces the mean error criterion MSE, which is used as an indicator of the differences between the expected values. The actual values are in statistical estimates. Simply put, the larger the sample size, the more accurate the statistical predictions are and the lower the standard deviation from the actual values, indicating that larger samples often contribute to improving the accuracy of statistical. estimates.

11. Recommendations

1. Using new types of survival regression models not mentioned in the research.
2. Use other estimation methods such as (weighted least squares method, moments, etc.) in estimating the parameters of survival regression models and comparing them with the Bayesian method used in the research.
3. Applying the survival regression models to real data.

reference

1. Abdel Moneim, Tharwat (2012) "*Using Bayesian statistical inference in the field of life tests*" University of Dammam.
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. Box, G. E. and Tiao, (1979)."*Distribution of Residual Autocorrelation in Multiple Autoregressive*", JASA, Vol. (74), pp. (928-934).
5. Collett, D. (2023). "*Modelling survival data in medical research"*.4th Edition CRC press.
6. Fox, J., & Weisberg, S. (2002). "*Cox proportional-hazards regression for survival data*". An R and S-PLUS companion to applied regression.
7. Haerting, J., Mansmann, U., & Chateau, L. (2007). " *Frailty Models in Survival Analysis*", Unpublished doctoral dissertation, Martin-Luther-Universidad Halle-Wittenberg.
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, Atheer Abdel Zahra (2018), "*Analysis of the survival function when the risk factor is proportional to time (an applied study)*" Master's thesis, College of Administration and Economics, University of Karbala.
12. Karim, M. R., & Islam, M. A. (2019). "*Reliability and survival analysis*". Springer Singapore.
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. 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.
15. Swami Nathan, H., & Gifford, J. A. (1982). "*Bayesian estimation in the Rash model*". Journal of Educational Statistics, 7(3), 175-191.