# Parametric Frailty Model with Time-Dependent Covariates

**Mohd Asrul Affendi Abdullah, Emir Mukhriz Zaimi, Siti Afiqah Muhamad Jamil**

***Abstract:*** ***The parametric frailty model has been used in this study where, the term frailty is used to represent an unobservable random effect shared by subjects with similar (unmeasured) risks in the analysis of mortality rate. In real-life environment, the application of frailty models have been widely used by biostatistician, economists and epidemiologist to donate proneness to disease, accidents and other events because there are persistent differences in susceptibility among individuals. When heterogeneity is ignored in a study of survival analysis the result will produce an incorrect estimation of parameters and standard errors. This study used gamma and Weibull distribution for the frailty model. The first objective of this study is to investigate parametric model with time dependent covariates on frailty model. The derivation is using either classical maximum likelihood or Monte Carlo integration. The second objective is to measure the effectiveness of Gamma and Weibull frailty model with and without time-dependent covariates. This is done by calculating the root mean square error (RMSE). The last objective is to assess the goodness of fit of Gamma and Weibull frailty model with and without time-dependent covariates using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Simulation is used in order to obtain the RMSE, AIC ad BIC value if time-dependent covariate does not exists. Between both models with time-dependent covariate, Weibull frailty distribution has lower AIC and BIC compared to Gamma frailty distribution. Therefore, Weibull frailty distribution with time-dependent covariate is preferable when a time-dependent covariate exists in a data.***

***Keywords*****: *Weibull Frailty, Gamma Frailty, AIC, BIC, RMSE***

## I. INTRODUCTION

The term frailty is introduced by [1] to indicate that every individual have different risks even though their measuring elements or attributes are quite similar or same. In survival analysis, the term frailty is used to represent an unobservable random effect shared by subjects with similar (unmeasured) risks in the analysis of mortality rate. For instance, there may be an association in the times to events of cancer diseases between siblings or married couples, even occurrence of nonlethal diseases within the same individual. Therefore, it is introduced in statistics in order to account the unmeasured risks.

In real-life environment, the application of frailty models have been widely used by biostatisticians, economists and epidemiologists to denote proneness to disease, accidents and other events because there are persistent differences in susceptibility among individuals.

Frailty is a measure of relative risk because the greater an individual's frailty, the greater the individual's tendency to the cause of death. According to Vaupel, Manton, and Stallard [1], this definition of frailty assumes that each individual is born at a level of relative frailty and stays at this level all its life. Usually, individuals within the same cluster share the same unobserved frailty [2].

There are many distributions that can be applied to frailty model. Instead of modelling using standard random effects models, censoring and truncation have been a problem to develop using inferential methods. [3] suggested that gamma distribution is the most widely used as frailty distribution since it is tractable. This is supported by [4] which use gamma and inverse Gaussian distribution and stated that both two distribution are relevant and mathematically tractable as a frailty distribution for heterogeneous populations. Other important information about frailty models is the shape of the distribution itself. The tails can determine the type of dependence of a frailty model. According to [5] if a certain distribution with a large right tail such as positive stable distribution leads to strong early dependence whereas, a distribution with a large left tail such as gamma and Weibull distribution lead to strong late dependence.

Covariate structure is assumed to act multiplicatively on a baseline hazard. The problem of ignoring frailty is investigated by [6]. In their analysis of event history data, there is no covariates effect on the response variable. Subsequently, TVC also applied in fixed covariate of right censored lung cancer data [7].

Consequently, the aims of study were to apply the parametric model with TVC on the frailty model and measure the effectiveness of Gamma and Weibull frailty based on the root mean square error (RMSE) and the model selection criteria of AIC and BIC. This study used R statistical software to complete the analysis.

## II. MATERIAL AND METHOD OF STUDY

### A. Framework of analysis

Figure 1 shows the flow of analysis which involves gamma and Weibull distribution of frailty model.

**Revised Manuscript Received on January 15, 2020**

***** Correspondence Author

**Mohd Asrul Affendi Abdullah***, Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Pagoh Campus, Johor, Malaysia.

**Emir Mukhriz Zaimi**, Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Pagoh Campus, Johor, Malaysia. **Siti Afiqah Muhamad Jamil**, Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Pagoh Campus, Johor, Malaysia.

This study uses the time-dependent covariate with application towards simulation data. There are two advantages in using gamma distribution which are the frailty distribution of survivors at any given age having the same values of parameter and different scale of parameter. The second advantage is that the frailty distribution among the persons dying at any given age having same shape parameter plus one, and a scale parameter as a function of the age at death.

Besides, for Weibull distribution, this frailty model is not as convenient as gamma because, the frailties are not be able to be integrated out in the conditional likelihood function leading to a non-tractable integral which needs to be approximated.

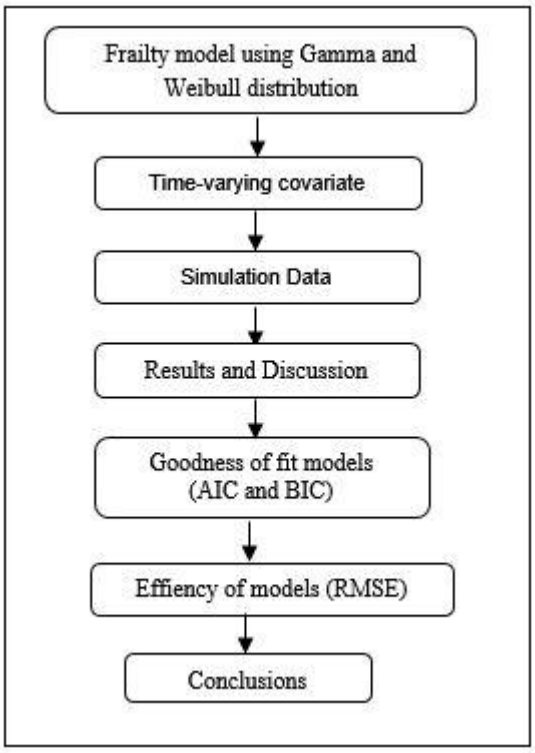


**Fig. 1. The framework of analysis**

Based on Figure 1 above, this study begins with the two parametric survival analysis which are the Gamma and Weibull model which then involve the frailty effect towards the data or covariates. By considering the time-varying effect, simulation study has been performed to observe the differences between the two frailty models. The AIC, BIC and RMSE has been used to compare the models.

### B. Gamma Frailty

In gamma frailty models, the restriction $\alpha = \lambda$ is used, which results in expectation of 1. The variance of the frailty variable is $\frac{1}{\lambda}$. Assume that the frailty term $u$ is distributed as gamma with $E(U) = 1$ and $Var(U) = \theta$. Then $\lambda = \alpha = 1/\theta$. The distribution function of the frailty term $u$ is then one-parameter gamma distribution similar as stated by [8].

$$g(u) = \frac{u^{\left(\frac{1}{\theta}-1\right)} e^{\left(\frac{-u}{\theta}\right)}}{\Gamma\left(\frac{1}{\theta}\right)\theta^{1/\theta}}, \theta > 0$$

$u > 1$ indicate that individuals in group $i$ are frail, whereas $u < 1$ indicate that individuals are strong and have lower risk.

### C. Weibull Frailty

Consequently, for Weibull frailty model, according to [4], the distribution of the frailty term $u$ can be written can be written as,

$$g(u) = \exp\left[-\left(\frac{u}{1/\Gamma\left(1+1/\eta\right)}\right)^{\eta} \cdot \left(\frac{\eta u^{(\eta-1)}}{\left(1/\Gamma\left(1+1/\eta\right)\right)^{\eta}}\right)\right]$$

$$= \exp\left\{-\left(u\Gamma\left(1+1/\eta\right)\right)^{\eta}\right\}\left(\eta u^{(\eta-1)}\right)\left(\Gamma\left(1+1/\eta\right)\right)^{\eta}$$

where $\eta = \phi - 1(\theta)$.

Though Weibull distribution is widely used in survival analysis, its marginal likelihood function is impossible to derive n an explicit form in frailty models. In order to perform parameter estimation, more sophisticated estimation strategies are required. For the gamma frailty model, we obtained the parameter estimates directly from the exact log-likelihood function using the classical maximum likelihood method. Unfortunately, we cannot use the same method for the Weibull distribution due the complexity of the likelihood function. The expectation and variance of a frailty term u following a Weibull distribution are [4],

$$E(U) = \alpha\Gamma\left(1+\frac{1}{\eta}\right) = 1$$

$$Var(U) = \alpha^2\left[\Gamma\left(1+\frac{2}{\eta}\right) - \left(\Gamma\left(1+\frac{1}{\eta}\right)\right)^2\right] = \theta$$

It follows that $\alpha = \frac{1}{\Gamma\left(1+\frac{1}{\eta}\right)}$ and $\theta = \frac{\Gamma\left(1+\frac{2}{\eta}\right)}{\left(\Gamma\left(1+\frac{1}{\eta}\right)\right)^2} - 1$,

where $\alpha$ and $\theta$ are functions of $\eta$. If the frailty $u$ follows a Weibull distribution, then the probability distribution function can be written as,

$$g(u) = e^{-\left(u\Gamma\left(1+\frac{1}{\eta}\right)\right)^{\eta}} \eta u^{(\eta-1)} \left(\Gamma\left(1+1/\eta\right)\right)^{\eta}, u > 0, \eta > 0$$

The marginal likelihood function is then given by, [11]

$$l(\psi,\eta,\theta,\beta)=\prod_{j=1}^{n_i}\int_0^{\infty}\left(h_0(t)ue^{\beta' z_{ij}}\right)^{\delta_{ij}}e^{-h_0(t)ue^{\beta' z_{ij}}}$$
$$\times e^{\left(-\left(u\Gamma\left(1+1/\eta\right)\right)^{\eta}\right)}\eta u^{(\eta-1)}\left(\Gamma\left(1+1/\eta\right)\right)^{\eta}du$$
$$=\prod_{j=1}^{n_i}\left(h_0(t)ue^{\beta' z_{ij}}\right)^{\delta_{ij}}\left(\Gamma\left(1+1/\eta\right)\right)^{\eta}$$

$$\int_0^{\infty}u^{d_i}e^{-\left(u\Gamma\left(1+1/\eta\right)\right)^{\eta}}\eta u^{(\eta-1)}e^{-\sum_{j=1}^{n_i}H_0(t)ue^{\beta' z_{ij}}}du$$

where $\psi$ is the vector of parameters of the baseline hazard function. Since the expression under the integral is not tractable, we propose to use Monte Carlo integration method to approximate the integral. The integration part of the likelihood function is,

$$\int_0^{\infty}u^{d_i}e^{-\left(u\Gamma\left(1+1/\eta\right)\right)^{\eta}}\eta u^{(\eta-1)}e^{-\sum_{j=1}^{n_i}H_0(t)ue^{\beta' z_{ij}}}du$$
$$=\int_0^{\infty}u^{d_i}e^{-\left(u\Gamma\left(1+1/\eta\right)\right)^{\eta}}e^{-\sum_{j=1}^{n_i}H_0(t)ue^{\beta' z_{ij}}}du$$

Using the transformation, $v=\sum_{j=1}^{n_i}H_0(t)ue^{\beta' z_{ij}}$ where, $dv=\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}du$ the likelihood function can be written as,

$$\int_0^{\infty}\left(\frac{u}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)^{d_i+\eta-1}\eta\exp\left[-\left(\frac{v\Gamma\left(1+1/\eta\right)}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)^{\eta}\right]$$
$$\times e^{-v}\left(\frac{dv}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)$$
$$=\int_0^{\infty}\frac{\Gamma(d_i+\eta)}{\left(\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}\right)^{d_i+\eta}}\eta\exp\left[-\left(\frac{v\Gamma\left(1+1/\eta\right)}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)^{\eta}\right]$$
$$\times\left(\frac{1}{\Gamma(d_i+\eta)}\right)\left(e^{-v}e^{d_i+\eta-1}\right)dv$$

Where $\left(\frac{1}{\Gamma(d_i+\eta)}\right)\left(e^{-v}e^{d_i+\eta-1}\right)$ is the probability distribution function of $\Gamma(d_i+\eta,1)$. It follows that the integration that is required is the expectation of,

$$\frac{\Gamma(d_i+\eta)}{\left(\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}\right)^{d_i+\eta}}\eta\exp\left[-\left(\frac{v\Gamma\left(1+1/\eta\right)}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)^{\eta}\right]$$

Evaluating this expectation is difficult. However, if we can draw sufficiently large $N$ samples from $\Gamma(d_i+\eta)$, we can estimate the expectation by,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\Gamma(d_i+\eta)}{\left(\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}\right)^{d_i+\eta}}\cdot\eta\exp\left[-\left(\frac{v\Gamma\left(1+1/\eta\right)}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)^{\eta}\right]$$

This is referred to as Monte Carlo integration. By Strong Law of Large Numbers, as $N\to\infty$ the estimated expectation will approach the true value.

The approximate marginal log-likelihood function is then given by,

$$\hat{l}(\psi,\eta,\theta,\beta)=\log(L(\eta,\eta,\theta,\beta))$$
$$=\sum_{j=1}^{G}\left[\left(\sum_{j=1}^{n_i}\delta_{ij}\left[\log\left(h_0(t)\right)+\beta' z_{ij}\right]\right)\right]$$
$$+\eta\log\Gamma\left(1+1/\eta\right)+\log\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\Gamma(d_i+\eta}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)\right)$$
$$\times\eta\exp\left[-\left(\frac{v_i\Gamma\left(1+1/\eta\right)}{\sum_{j=1}^{n_i}H_0(t)e^{\beta' z_{ij}}}\right)^{\eta}\right]$$

where $v_1$, ..., $v_N$ are $N$ independent realizations of $\Gamma(d_i+\eta,1)$. In order to achieve accurate estimates, $N$ is required to be greater than 10,000 [9].

If computer time is an issue, we also propose to use quadrature method to approximate the integral in Equation (3.53). This method is available in statistical packages such as R, SAS and S-PLUS.

$\Psi=(\lambda)$ for exponential baseline hazard and $\psi=(\eta^*,\alpha^*)$ for the Weibull baseline hazard. The maximum likelihood estimates of $\beta$ and $\eta$ can be obtained by maximizing the loglikelihood function. Once we obtain the estimate of $\eta$, the frailty variance $\theta$ can be estimated by,

$$\theta = \frac{\Gamma\left(1+\frac{2}{\eta}\right)}{\left(\Gamma\left(1+\frac{1}{\eta}\right)\right)^2} - 1$$

The first-order derivatives can be approximated with simulations in order to obtain the maximum-likelihood estimates. $\eta^*$ and $\alpha^*$ are not of interest and hence we will treat them as fixed values in our simulation study. Note that the baseline hazard parameters for the Weibull distribution should be distinguished from the lifetime frailty parameters, $\eta$ and $\alpha$.

### D. Model Selection

For simulation study, the number of sample size is repeated for 1000 times with 100 number of sample size. The data is simulated for all models and analysed based on their AIC, BIC and RMSE values.

Additionally, study compared the AIC and BIC to select the best fit model towards simulations data of the two models of gamma and Weibull distribution.

Where $k$ is the number of parameters while $n$ is the size of sample. The smaller value of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) is better while the largest value of likelihood function is the best fit model with formula given by:

$$AIC = -2\log(likelihood) + 2(k)$$

$$BIC = -2\log(likelihood) + k\log(n)$$

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

Before that, in order to be certain on the simulation procedure, study have to compare the values of estimated value of the coefficient and the estimated coefficient. As the results show that the simulated data have a quite close value to the real one, then the data is correctly estimated. Besides, coverage probability is also used to make sure that the estimated data lies within the range.

The efficiency of the models can be observed based on the value of root mean square error (RMSE). This study uses R software to complete the analysis.

## III. RESULTS AND DISCUSSION

First of all, study involve three variables of simulation data which are the age, gender and treatment that is based on rectum cancer of patient data. Table below shows the detail about the simulated data by using R software.

**Table 1 Summary of Data**

| Variable | Coding |
|---|---|
| Age | rnorm(n) |
| Gender | rnorm(n) |
| Treatment | rnorm(n) |
| Frailty | rgamma(n,shape-1, rate=200) |
| Time-dependent covariate | rgamma(n,shape-1, rate=200) |

In order to observe the true values of covariates towards the model, study calculate the coverage probability so that the proportion time of true value could be estimated as it contains true value within an interval. By referring to [10] if the coverage probability is greater than 70%, the data is simulated within the range.

Table 2 below shows the summary of the coverage probability between models with time-dependent covariates and fixed covariates. Model 1, 2, 3 and model 4 are based on the following:

Model 1: Gamma Frailty Model
Model 2: Weibull Frailty Model
Model 3: Gamma Frailty with time-dependent covariates
Model 4: Weibull Frailty with time-dependent covariates

Before that, by comparing the results of coefficient and the estimated coefficient, it is not approximately estimated to the true value. Then, study have to observe the value of coverage probability to proceed with the analysis. Based on Table 2, most of the result indicates that the estimated probability is above 70%. Thus the data is simulated correctly and accepted as the estimated value not too far from the actual value.

**Table 2 Summary of Coverage Probability.**

| Models | X1 | X2 | X3 | Z | TV |
|---|---|---|---|---|---|
| Model 1 | 1.00 | 1.00 | 0.99 | 0.96 | - |
| Model 2 | 0.99 | 0.97 | 0.99 | 0.97 | - |
| Model 3 | 0.97 | - | - | 0.98 | 0.97 |
| Model 4 | 0.98 | - | - | 0.97 | 0.98 |

Besides, in order to check for the efficiency of the model, study observe the root mean square error (RMSE). The lower RMSE indicates the model is more efficient.

**Table 3 Summary of RMSE**

| Models | RMSE |
|---|---|
| Model 1 | 8.1938 |
| **Model 2** | **4.6173** |
| Model 3 | 5.6292 |
| **Model 4** | **3.3266** |

According to the Table 3 above, study summarised the value of RMSE based on model 1, 2 3 and model 4. From the results, Weibull frailty model with time-dependent covariates shows the lowest RMSE compared to other models. This indicates that the data fits well in the model and has the highest measure of accuracy compared to others. The table also shows that the model with

time-dependent covariate shows lower RMSE compared to the model without time-dependent covariate.

In order to choose the best fit model, study applied the goodness of fit test for the four models by comparing the values of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Table 4 below shows the output from the simulation procedure of AIC and BIC value.

**Table 4 Summary of Goodness-of-fit**

| Models | AIC | BIC |
|---|---|---|
| **Model 1** | **234.93255** | **245.35323** |
| Model 2 | 724.52991 | 734.95059 |
| Model 3 | 597.5076 | 609.9283 |
| **Model 4** | **596.1812** | **608.6018** |

The AIC and BIC values which are the smallest is preferred in determining which model has the best goodness-of-fit. Based on the table above, gamma frailty model shows the lowest value compared to other models. However, when time-dependent covariate exists, the value of AIC and BIC increase. Between both on the models with time-dependent covariates, Weibull shows a lower value compared to the other models.

## IV. CONCLUSION

The simulation is repeated for 1000 times with 100 number of sample size. The model with the lowest value would be chosen as the best fit model. The results have been divided into two different conclusions.

Overall, when comparing the Gamma and Weibull frailty based on the simulation study, the smallest value of AIC, BIC and the RMSE are based on Model 4 which consider Weibull frailty with time-dependent covariate to be the best fit model.

Besides, the results also the lowest at model 1 for the AIC and BIC value while the lowest at model 2 for RMSE which is the Weibull and Gamma frailty model without time varying covariate.

In short, the simulation data with corresponding variable fit towards both frailty models which are the Weibull and Gamma while for the corresponding frailty model with time varying covariate would fit only for the Weibull frailty model (model 4).

## V. RECOMMENDATION

For future research, real data simulation using frailty models with time-dependent covariates is advised to get more accurate results. Researcher also may use other method to provide some illustrative example for the real-life data.

Other improvement that can be made is by comparing multivariate frailty models with time dependent covariates and univariate frailty models with time-dependent covariates. The work will not be too complicated because multivariate frailty model is actually an extension of the univariate frailty model that allows individuals in the same cluster to share the same frailty value.

## ACKNOWLEDGEMENT

We would like to thank the Ministry of Higher Education and Universiti Tun Hussien Onn Malaysia, for supporting this research project for fundamental research grant (FRGS Vot K178)

## REFERENCES

1. Vaupel, J. W., Manton, K. G., and Stallard, E. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16: 439-454.
2. Shih, J. H. and Louis, T. A. 1995. Assessing gamma frailty models for clustered failure time data. Lifetime Data Analysis 1, 2: 205-220.
3. Clayton, D. G. 1978. A model for association in bivariate life tables and its application in epidemiological studies in familial tendency in chronic disease incidence. Biometrika, 65: 141-151.
4. Hougaard, P. 1986. Survival models for heterogeneous populations derived from stable distributions. Biometrika, 73: 387-396.
5. Hougaard, P. 2000. Analysis of Multivariate Survival Data. New York: Springer.
6. Flinn. C. J. and Heckman, J. J. 1982. New methods for analysing individual event histories. In: Sociological Methodology p. 99-140.
7. Jamil, S. A. M., Abdullah, M. A. A., Kek, S. L., Olaniran, O. R., & Amran, S. E. (2017, September). Simulation of parametric model towards the fixed covariate of right censored lung cancer data. In Journal of Physics: Conference Series (Vol. 890, No. 1, p. 012172). IOP Publishing.
8. Rondeau, V., Commenges, D., and Joly, P. 2003. Maximum penalized likelihood estimation in a gamma-frailty model. Lifetime Data Analysis, 9(2), 139-153.
9. Balakrishnan, N. and Peng, Y. (2006). Generalized gamma frailty model. Statistics in Medicine, 25: 2797-281.
10. Kysely, J. 2009. Coverage probability of bootstrap confidence intervals in heavy tailed frequency models, with application to precipitation data. Theo. Appl. Climatol, 101: 345-361.
11. Klien, J. P. 1992, Semi-parametric estimation of random effect using Cox model based on the EM algorithm. Biometrics, 48: 795-806.

## AUTHORS PROFILE

**Mohd Asrul Affendi Abdullah,** Ph.D. in Biostatistics at Universiti Sains Malaysia, Health Campus, Member of International Society for Clinical Biostatistics, Member of Science and Mathematics Malaysia Society. Research in Biostatistics, Survival Analysis, Inflated Model, Environmental Statistics and High Dimensional Regression. Senior Lecturer at Universiti Tun Hussein Onn Malaysia.

**Emir Mukhriz Zaimi,** M.Sc in Industrial Statistics at Universiti Tun Hussein Onn Malaysia. Research work in Frailty Survival Analysis and Environmental Statistics. Operational Analyst at Pos Malaysia Berhad.

**Siti Afiqah Muhamad Jamil**, Ph.D. student at Universiti Tun Hussein Onn Malaysia. Member of Malaysia Institute of Statistics (ISM). Research work in Survival Analysis, Generalized Linear Model, Logistic, Linear and Multivariate Regression Analysis, Environmental Statistics and Biostatistics.