www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 11 November, 2014 Page No. 9172-9177

Gamma Distributions Model For The Breast Cancer Survival Data Using Maximum Likelihood Method

***K. H. Khan*[1]*, M. Saleem*[2]**

[1]Department of Mathematics,College of Science and Humanities,Salman Bin Abdulaziz University, Al-Kharj,
Kingdom of Saudi Arabia
Email: drkhizar@gmail.com
&
[2]Centre for Advanced Studies in Pure and Applied Mathematics, B.Z. University Multan, Pakistan.
E-mail: colsaleem_2009@yahoo.com

Abstract: ***The breast cancer censored data of 254 patients was considered for the survival rate estimates. The data [12, 18] was treated at the chemotherapy department, Bradford Royal Infirmary for ten years. Here in this paper Gamma probabilitydistribution model is used to obtain the survival rates of the patients (see [2], [6], [13]). Maximum likelihood method has been used through unconstrained BFGS optimization method [5, 8, 9, 10] (BFGS-Broyden Fletcher-Goldfarb and Shanno Method) to find the parameter estimates and variance-covariance matrix for the Gamma distribution model. Finally the survivor rate estimates for the parametric Gamma probability model has been compared with the non-parametric (Kaplan-Meier-[15]) method.***

Keywords: Gamma distribution model, Censoring, Breast Cancer Data sets, BFGS-unconstrained optimization method, Maximum likelihood function and Kaplan-Meier survivor rate estimates.

1. Introduction

Breast cancer is a systemic disease (see [4], [12],[18])until proved otherwise. When the treatment is stopped the disease progresses with uniform ‘velocity’ *v* through a fixed ‘distance’d in the disease to recurrence point.

In this paper, we find the parameter estimates, survival rate estimates, variance covariance matrixfor the Gamma probability distribution model using maximum likelihood function using breast cancer data [12].

For the survival of the patient with the breast cancer, a statistical approach is considered; wihich is based on two parameters refered as scale and shape parameters respectively of the said distributions. Further work on probabilistic approach has been done by Khan, K.H. [16]. using Inverse Guassioan distribution model. The survivor rate estimates for the Gamma probability distribution has also been compared with the non-parametric model [15].

2. *The Gamma Model and Estimation of Parameters*

The data regarding survival analysis generally falls in two classes: (i) the failure time of items, which actually fail during the experiment, (ii) the survival times of items which, actually survive with the experiment.

These classes are generally separated statistically by the use of censoring, for detail see Cox, [8]. In parametric models the *pdf* of lifetime ‘T’ has form $f(t,\theta)$ with survival function$R(t,\theta)=P(T>t)$, where θ is a vector of parameters. The contribution to the likelihood of an item that fails at time *t* is $f(t,\theta)$ and an item that survives beyond time is$R(t,\theta)$. Thus, according to the Lawless [16], using the Gamma distribution models, the likelihood function when the time is divided into intervals is given as

$$L=\prod_{i=1}^{NG}\left[F(t_i)-F(t_{i-1})\right]^{f_i}\left(1-F(T_n)\right)^{N-F},$$

whereNG, f_i, N and F are the number of recurrence groups, number of failures (recurrences) in the *i*th year, sample size and total number of recurrences in 10 years respectively.

The maximum likelihood estimates can be obtained by takingthe log-likelihood function. Since the probability of no failure until time t is defined by$R(t)=1-F(t)$, then the log-likelihood function ($\ln L$)can be written as

$$\ln L \;=\; \sum_{i=1}^{10} f_i\left[\ln\left(R(t_{i-1}) \;-\; R(t_i)\right)\right] + \;(N-F)\;\ln\left(R(T_n)\right)$$

(2.2)

To find the parameter estimates we used the unconstrained optimization method '(BFGS-Broyden Fletcher-Goldfarb and Shanno Method (see. [8], [9]). The BFGS - Quasi-Newton-Method is an iterative method, which minimizes the objective functionand requires only first partial derivatives in addition to the function values. So, the log-likelihood function to be maximized is equivalent to the minus times the log-likelihood function to be minimized. Therefore the required form for the estimation parameters is ($\ell = -\ln \mathrm{L}$). The variance-covariance matrix of estimates $\hat{a}$ and $\hat{b}$

$$\begin{bmatrix} \dfrac{\partial^2 \ell}{\partial a^2} & \dfrac{\partial^2 \ell}{\partial a \partial b} \\ \dfrac{\partial^2 \ell}{\partial a \partial b} & \dfrac{\partial^2 \ell}{\partial b^2} \end{bmatrix}^{-1}$$

is calculated automatically and numerically as a part of these optimization procedures, and without any direct evaluation of the second derivatives of ℓ which would be very complicated.

The Gamma distribution is extensively used in engineering, reliability, and applied statistics. Gupta and Groll (1961) discussed the use of the gamma distribution in acceptance sampling based on life tests. Johnson and Kotz (1970) have given a good general review of the gamma distribution. The gamma distribution has also received considerable attention in the area of weather analysis.

The two-parameter gamma distribution of a random variable T has a pdf of the form

$$f(t;\lambda,\alpha) = \frac{\lambda^{\alpha} t^{\alpha-1} e^{-\lambda t}}{\Gamma_{(\alpha)}}, t \geq 0$$

(2.3)

where $\lambda > 0$, and $\alpha > 0$ are the scale and shape parameters of the gamma density function respectively, and

$$\Gamma_{(\alpha)} = \int_0^{\infty} v^{\alpha-1} e^{-v} dv.$$

From the above gamma density function, if we take $\alpha = 1$ the gamma density function reduces to exponential death density function. If $\alpha > 1(<1)$, then the failure rate or hazard rate of the gamma density function increases (decreases) as a function of time.

For ease of computation, we take $\lambda = \dfrac{1}{\theta}$ so the gamma death density function reduces to

$$f(t;\theta,\alpha) = \frac{t^{\alpha-1} e^{-\frac{t}{\theta}}}{\theta^{\alpha}\Gamma_{(\alpha)}}, t \geq 0, \theta, \alpha > 0 .$$

Now, the gamma survival distribution is given by

$$S(t) = \int_t^{\infty} f(x) dx$$

$$\Rightarrow S(t) = \int_{\frac{t}{\theta}}^{\infty} v^{\alpha-1} e^{-v} dv$$

$$\Rightarrow S(t) = 1 - I\left(\alpha, \frac{t}{\theta}\right) = Q\left(\alpha, \frac{t}{\theta}\right),$$

where

$$I\left(\alpha, \frac{t}{\theta}\right) = \frac{1}{\Gamma_{(\alpha)}} \int_0^{\frac{t}{\theta}} v^{\alpha-1} e^{-v} dv$$

is called incomplete gamma function.

Now the hazard rate is

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\theta}\left(\frac{t}{\theta}\right)^{\alpha-1} e^{-\frac{t}{\theta}}}{Q\left(\alpha, \frac{t}{\theta}\right)}, t \geq 0 . \qquad (2.8)$$

Again the modified likelihood function for the failed and censored times is given by

$$\ell = -\sum_f \ln(f(t_i;\underline{\theta})) - \sum_c \ln(S(t_i;\underline{\theta})) ,$$

where first sum is over failures and the second is over censored items. Hence

$$\ell = \sum_u \left[\alpha \ln\theta + \ln\Gamma_{(\alpha)} - (\alpha-1)\ln t_i + \frac{t_i}{\theta}\right] + \sum_c \ln\left[Q\left(\alpha, \frac{t_i}{\theta}\right)\right]$$

Now the partial derivatives are given by

$$\frac{\partial \ell}{\partial \alpha} = \sum_u \left[\ln\theta + \psi(\alpha) - \ln t_i + \frac{t_i}{\theta}\right] + \sum_c \left[\frac{\partial Q\left(\alpha, \frac{t_i}{\theta}\right) \big/ \partial\alpha}{Q\left(\alpha, \frac{t_i}{\theta}\right)}\right]$$

(2.11)

(2.4)

$$\frac{\partial \ell}{\partial \theta} = \sum_u \left[\frac{\alpha}{\theta} - \frac{t_i}{\theta^2}\right] - \sum_c \left[\frac{\partial Q\left(\alpha, \frac{t_i}{\theta}\right) \big/ \partial\theta}{Q\left(\alpha, \frac{t_i}{\theta}\right)}\right],$$

where $\psi(\alpha) = \dfrac{d}{d\alpha}\left(\ln(\Gamma_{(\alpha)})\right)$ is called the Psi Function.

To find the parameter estimates, we took benefit of the algorithm of Moor (1982) in which the incomplete gamma integral of eq. (2.7) was computed with its first derivatives w.r.t. αand θ. The algorithm of Moor (1982) itself uses the algorithm of Bhattacharjee (1970) for finding the incomplete gamma integral.

Harter and Moor (1965) considered three-parameter gamma distribution and applied the maximum likelihood principle to find the parameter estimates. We have noted that exponential distribution is a special case of the gamma distribution for $\alpha = 1$. Several authors have considered the problem of estimating the parameters of the Gamma distribution (see [2, 6, 13]).

The solutions of eq. (2.11) and eq. (2.12) yields the parameter estimates, $\hat{\alpha}, \hat{\theta}$ using numerical optimization techniques. The subroutine was used the BFGS unconstrained optimization techniques to find the parameter estimates, the variance-covariance matrices, survivor rate estimates and maximum likelihood function for the Gamma distribution.

3. *APPLICATION*

We considered the data of 254 patients surviving with breast cancer. These patients were initially treated at the department of chemotherapy department, Bradford Royal Infirmary, [12], England, thirty five years ago. Each patient was treated for a period of ten years or until death. The patients surviving with breast cancer were between 23 and 82 years old (Hancock et al. [12]).The patients were classified into four diffrenet stages using TNM (Tumor Nodes Metastases) system and clinically staged accordingly.

Out of 254 patients, 100 patients were premenopausal and 154 were postmenopausal. A woman was considered to be postmenopausal when 2 years had elapsed since her last menstrual period. The two main categories are premenopausal and postmenopausal. Note that Stages I & II for premenopausal and postmenopausal were each combined together.

Table-1. Age Distribution Related to Clinical Stage and Menopausal Status

Patient	Stage I		Stage II		Stage III		Stage IV	
Age	Pre-	Post-	Pre-	Post-	Pre-	Post-	Pre-	Post-
21-30	-	-	-	-	2	-	1	-
31-40	6	-	1	-	12	-	11	-
41-50	16	4	8	2	17	3	16	7
51-60	1	13	-	3	5	29	4	16
61-70	-	12	-	1	-	27	-	24
71-80	-	3	-	1	-	4	-	4
81-90	-	-	-	1	-	-	-	-

Table-2. Survivalsand Failures Related to Clinical Stage and Menopausal Status

Stage	Menopausal Status	Surviving with Cancer	Surviving with Recurren	Dying without	Dying with Recurren	Dying with Canc	Patients in each
Stage I	Pre-	16	4	1	0	2	23
	Post-	8	5	4	1	14	32
Stage II	Pre-	5	1	0	0	3	9
	Post-	1	0	1	1	5	8
Stage III	Pre-	6	2	1	1	26	36
	Post-	5	4	2	7	45	63
Stage IV	Pre-	0	0	0	0	32	32
	Post-	1	1	0	3	46	51

Table-3. Data for Stages I to IVover the ten years

Time (Years)	Stage-I &II Pre-menopausal		Stage-I & II Post-menopausal		Stage-III Pre-menopausal		Stage-III Post-menopausal		Stage-IV Pre-menopausal		Stage-IV Post-menopausal	
	Survivers	Failures	Survivers	Failures	Survivers	Failures	Survivers	Failures	Survivers	Failures	Survivers	Failures
0	32	0	40	0	36	0	63	0	32	0	51	0
1	32	0	38	2	30	6	58	5	23	9	35	16
2	31	1	35	3	24	6	53	5	13	10	22	13
3	31	0	35	0	22	2	43	10	4	9	11	11
4	30	1	32	3	18	4	36	7	2	2	7	4
5	29	1	27	5	15	3	30	6	2	0	4	3
6	29	0	25	2	14	1	22	8	1	1	4	0
7	28	1	22	3	12	2	19	3	1	0	4	0
8	27	1	20	2	12	0	17	2	1	0	2	2
9	27	0	18	2	9	3	13	4	1	0	2	0
10	26	1	14	4	8	1	9	4	0	1	2	0

Table-4 Estimates of Parameters and ML-Function for GumbelDistribution Model

Estimates	Pre-menopausal			Post-menopausal		
	Satge-I&II	Satge-III	Stage-IV	Satge-I&II	Satge-III	Stage-IV
$\hat{a}$	0.354132	0.37343	0.35135	0.36467	0.214083	0.32153
$\hat{b}$	0.02105	0.02953	0.01536	0.02435	0.03565	0.094536
MLF	4.53565	119.2536	115.89472	9.254892	125.25354	112.5691

Table-5.Estimates of Variance-Covariance Matrix and Gradient vector for the Gamma Model

Pre-Menopausal Stages			
Variance	Stage-I & II	Stage-III	Stage-IV

Covariance Matrix	0.0053253 -0.000558 -0.000558	0.002136 -0.000254 -0.000254	0.002587 -0.000245 -0.000245
Gradient	-0.2154E-07	-0.53466E-07	-0.8235E-09
Post-Menopausal Stages			
Variance	Stage-I&II	Stage-III	Stage-IV
Covariance Matrix	0.0012578 -0.0005421	0.0005755 -0.000587	0.000851 -0.000087
Gradient Vector	0.24102E-06 0.72365E-07	-0.51022E-05 -0.8213E-07	0.73206E-06-0.4216E-07

TABLE-6. Survival Proportion for Pre-menopausalStages

Time (Years)	Stage I & II		Stage III		Stage - IV	
	Kaplan-Mier	Gamma	Kaplan-Mier	Gamma	Kaplan-Mier	Gamma
1	1.00000	0.971421	0.833333	.78254	0.71875	0.55256
2	0.96875	0.965486	0.666666	.74569	0.40625	0.41565
3	0.96875	0.951057	0.611111	.71356	0.12500	0.19851
4	0.93750	0.941246	0.499999	.68456	0.06250	0.09856
5	0.90625	0.932845	0.416666	.65349	0.06250	0.01568
6	0.90625	0.924219	0.388888	.54635	0.03125	0.00654
7	0.87500	0.89598	0.333333	.48359	0.03125	0.00321
8	0.84375	0.884698	0.333333	.43569	0.03125	0.00123
9	0.84375	0.85387	0.249999	.39564	0.03125	0.00035
10	0.81250	0.82657	0.222222	.28654	0.00000	0.000013

TABLE-7. Survival Proportion for Post-menopausal Stages

Year	Stage I & II		Stage III		Stage - IV	
	Kaplan-Mier	Gamma	Kaplan-Mier	Gamma	Kaplan-Mier	Gamma
1	0.9500	0.92435	0.9206	0.85642	0.6862	0.52654
2	0.8750	0.90125	0.8412	0.81205	0.4313	0.41052
3	0.8750	0.88659	0.6825	0.75698	0.2156	0.38412
4	0.8000	0.82310	0.5714	0.71265	0.1372	0.31058
5	0.6750	0.79951	0.4761	0.67524	0.0784	0.21546
6	0.6250	0.73528	0.3492	0.58654	0.0784	0.16542
7	0.5500	0.61256	0.3015	0.46895	0.0784	0.05465
8	0.5000	0.56213	0.2698	0.31525	0.0391	0.02413
9	0.4500	0.46684	0.2063	0.25413	0.0391	0.008794
10	0.3500	0.336041	0.1428	0.19356	0.0391	0.002469

5. CONCLUSIONS

Analysis shows that the Gamma distribution is a reasonable model to describe the progression of breast cancer and finding survivor rate estimates for the medical data. Using Maximum likelihood method through unconstrained optimization method (BFGS- Broyden Fletcher-Goldfarb and Shanno Method [10]) the parameter estimates and variance-covariance matrix for the Gamma distribution modelpresented.However unlike a number of two-parameter distributions which are used in survivor studies it does have some beaming on the physical process being described.

REFERENCES

[1] Abramowitz, M. and Stegun, I. A. (1972). Handbook of Mathematical Functions. New York: Dover. Chapter 6: Gamma and Related Functions.

[2] Berman, M. (1981), “The maximum likelihood estimators of the parameters of the gamma distribution are alwayspositively biased”,Communications in Statistics, A, 10, 693-697.

[3] Bhattacharjee GP (1970). The incomplete gamma integral. Appl Statist 19:285-287.

[4] Boag, J.W. (1949). *Maximum Likelihood Estimations of the Population of Patients Cured by Cancer Therapy (With Discussion).* J.R. Stat. Soc. Series B, 1, No.11, pp.15 - 53.

[5] BroydenC.G., J.E. Dennis Jr., J.J. MoréOn the local and superlinear convergence of quasi-Newton methodsJ.

Inst. Math. Appl., 12 (1973), pp. 223–246.

[6] Choi, S.C. and Wette, R (1969) Maximum Likelihood Estimation of the Parameters of the Gamma Distribution

and Their Bias, Technometrics. 11, 4, pp. 683-690.

[7] Cox, D.R and Oaks, D. (1984). *Analysis of Survival Data. London: Chapman and Hall.*

[8] Davidon, W.C. (1959). O*ptimally conditioned optimization algorithms without line searches*, Mathematical Programming, 9, pp. 1-30.

[9] Dong-Hui Li, Masao Fukushima (2001).A modified BFGS method and its global convergence in nonconvex minimization, Journal of Computational and Applied Mathematics, Volume 129, Issues 1–2, pp. 15–35.

[10] Fletcher R. and M.J.D.Powell (1963). *Rapidly convergent descent method for minimization*, The Computer Journal, 6, pp. 163-168.

[11] Gupta, S. S. and Groll, P.A. (1961), Gamma distribution in acceptance sampling based on life tests, Journal of the American Statistical Association, vol. 56, 942 - 970

[12] Hancock, K. Peet, B. G., Price. J., Watson, G. W., Stone, J. and Turner, R. L. (1977). *Ten Year Survival Rate in Breast Cancer Using Combination Chemotherapy*, British Journal of Surgery, 64, pp.134 - 138.

[13] Harter H.Leon and Moore, Albert H. (1965). Maximum Likelihood estimation of the parameters of Gamma and Weibull populations from complete and from censored samples. Technometrics 7,639-643.

[14] Johnson, N. L., and Kotz, S. (1970), Continuous Univariate Distributions-I, New York: John Wiley.

[15] Kaplan, E.L., P. Meier, P. (1958). Nonparametric estimation from incomplete observations, J.Amer. Statist. Assoc., 53, pp. 457-481.

[16] Khan K.H., Zafar Mehmud (2002). Inverse Gaussian distribution Model for Cancer SurvivalData Using Maximum Likelihood Method Appl. Comput. Math, An International Journal, 1(2), 201-209.

[17] Moore, R. J. (1982). Algorithm AS 187: Derivatives of the incomplete gammaintegral. Appl. Statist., 32, 330-335.

[18] Watson, G. W. and Turner, R. L. (1959). *Breast cancer: A new approach to therapy.* British Medical Journal, 1, pp. 1315 - 1320.