IJST EXPLORE International Journal of Science and Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-7318

International Journal of Science and Technology Volume 3 No. 3, March, 2014

# Parametric Regression Models in the Analysis of Breast Cancer Survival Data

**[1]V. Vallinayagam, [2]S. Prathap, [3]P. Venkatesan**

[1]Department of Mathematics, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India.
[2]Department of Mathematics, Jeppiaar Engineering College, Chennai, Tamil Nadu, India.
[3]Department of Statistics, National Institute for Research in Tuberculosis,
Chennai, Tamil Nadu, India.

## ABSTRACT

Parametric regression models are widely used in the modeling of survival data under various diseases. The aim of this paper is to compare the performance of the common parametric models namely, Exponential, Weibull, Gompertz, Lognormal and Loglogistic using Breast Cancer data. The result shows that Lognormal model is better than other models.

**Keywords:** *Parametric Models, Lognormal, Deviance, Breast Cancer.*

## 1. INTRODUCTION

Regression modeling of the relationship between an outcome variable and one or more predictor variables is commonly employed in all fields. The popularity of this approach is due to the fact that plausible models may be easily fit, evaluated and interpreted. Statically, the specification of a model requires choosing both systematic and error components. The choice of the systematic components involves an assessment of the relationship among the "average" of the outcome variable relative to specific levels of the independent variables (Hosmer, Lemeshow and May, 2008).

The Accelerated Failure Time model which simply regresses the logarithm seldom been utilized in the analysis of censored survival data in clinical trial. This log-linear model has been studied extensively with observations subject to right censorship (Buckley and Jama 1979; Koul, Susaria and Van Ryzin 1981); Leurgans 1987; Miller 1976; Prentice 1978. The Accelerated Failure Time model performs better than the proportional hazard model in applications where the effects of treatment are to accelerate or delay the event of interest (Kay and Kinnersley, 2002). The main reason of the unpopularity of AFT model is its complicated estimation process, even if the data set consists of small number of covariates (Jin, Lin, Wei and Ying, 2003). Description of parametric survival models in detail (Coax and Oakes, 1984; Kalbfleisch and Prentice, 1980). Comparison of five survival models for breast cancer data (Hayat et al., 2010). Comparison of survival models for tuberculosis clinical trial data (Ponnuraja and Venkatesan, 2010). Comparison of distribution for HIV virus data in San Fransisco (Byers et al, 1988).

The objective of this paper is to compare the performance of parametric models using German Breast Cancer data.

## 2. DATA FOR STUDY AND DISTRIBUTION

We consider in this section, a real-life data set obtained from ftp//ftp.wiley.com/public/scitech_med/survival. This data consist of the German Breast Cancer study of 686 patients with 16 variables. The event of interest is survival time. These are the covariates considered here, 1. Age (Years), 2. Menopausal status
(1 = Yes and 2 = No), 3. Harmone Therapy (1 = Yes and 2 = No), 4. Tumor Size (mm), 5.Number of Nodes involved (1 - 51), 6. Tumor Grade (1 – 3), 7. Number of Progesterone Receptors (1 – 2380) and
8. Number of Estrogen Receptors (1 – 1144). Event is coded as 1 and censoring is coded as 0.

### 2.1. Exponential Distribution

A random variable T has the Exponential distribution with the following hazard, density and survivorship functions

$$h(t\,,\lambda) = \lambda$$

$$f(t,\lambda) = \lambda \exp(-\lambda t)$$

$$S(t,\lambda) = \exp(-\lambda t)$$

where $\lambda>0$.

The exponential distribution is also of practical importance because of its simplicity, and of theoretical importance because it is the break over point between increasing failure-rate models and decreasing-failure-rate models (Bain, 1964). Zelen and Dannemiller (1961) have demonstrated that the use of a one-parameter exponential model for failure analysis. The Best Estimate of Reliability in the Exponential case (Pugh, 1963).

### 2.2. Weibull Distribution

A random variable T has the Weibull distribution with the following hazard, density and survivorship functions

$$h(t\,,\lambda,\gamma) = \lambda\,\gamma t^{\gamma-1}$$

$$f(t,\lambda,\gamma) = \lambda\,\gamma t^{\gamma-1} \exp(-\lambda t^{\gamma})$$

$$S(t,\lambda,\gamma) = \exp(-\lambda t^{\gamma})$$

where λ>0, γ>0 .

Weibull distribution which is generalized version of the exponential distribution is widely used in modeling weather forecasts in meteorology, and defining the distribution of wind speed in radar modeling. Weibull distribution is preferred for performing survival data analysis in industrial engineering uses (Weibull, 1951). Weibull distribution is an important distribution model used in medicine since it is an flexible distribution that allows a monotonous increasing and decreasing of mortality ratio in patient groups. The distribution of the survival period of childhood leukemia patients was analysed using Weibull distribution (Viscomi et al., 2006). In a study conducted in Italy on the nationwide estimations were made for defining the parameters of the Weibull distributions (Inghelmann et al., 2005).

### 2.3. Gompertz Distribution

A random variable T has the Gompertz distribution with the following hazard, density and survivorship functions

$$h(t\ ,\lambda,\gamma) = \lambda \exp(\gamma t)$$

$$f(t,\lambda,\gamma) = \lambda \exp(\gamma t)\ exp\left(\frac{\lambda}{\gamma}\left(1 - exp(\gamma t)\right)\right)$$

$$S(t,\lambda,\gamma) = exp\left(\frac{\lambda}{\gamma}\left(1 - exp(\gamma t)\right)\right)$$

where λ>0, γ∈(-∞,∞).

Gompertz model used frequently by medical researchers and biologists in modeling the mortality ratio data, was formulated by Benjamin Gompertz in 1825. Gompertz is a growth model and has been used in relation with tumor development. The Gompertz distribution is often applied to describe the distribution of adult lifespans by demographers (Vaupel, James, 1986; Preston, Samuel et al., 2001) and actuaries (Benjamin et al., 1980; Willemse and Koppelaar, 2000), Related fields of science such as biology (Economos, 1982) and gerontology (Brown and Forbes, 1974) also considered the Gompertz distribution for the analysis of survival. More recently, computer scientists have also started to model the failure rates of computer codes by the Gompertz distribution (Ohishi, Okamura and Dohi, 2009). In marketing science, it has been used as an individual-level model of customer lifetime (Bemmaor, Albert and Glady, 2012). Ahuja and Nash (1976), showed that Gompertz distribution was, with a simple conversion, related to some distributions in the Pearson distribution family.

### 2.4. Lognormal Distribution

A random variable T has the Lognormal distribution with the following density, survivorship and hazard functions

$$f(t,\mu,\sigma^2) = \frac{\exp[-(logt - \mu)^2/2\sigma^2]}{\sqrt{2\pi}\sigma t}$$

$$S(t,\mu,\sigma^2) = 1 - \phi\left(\frac{logt - \mu}{\sigma}\right)$$

$$h(t\ ,\lambda,\gamma) = \frac{f(t)}{S(t)}$$

Where σ>0.

The theory of the lognormal distribution was characterized by Mc Alister in 1897. There is accordance to the lognormal distribution in many examples in area of medicine. The studies on determining the survival in cancer, the studies by Horner to determine the beginning age of Alzheimer's disease may be among the examples of the medical studies (Limpert, Stahel 2001; Horner, 1987). The limited-stage small-cell lung cancer patients with Kaplan-Meier curves, Cox proportional hazard model, Boag log-normal (cure rate model) and log-normal survival analysis methods analyzed (Tai et al., 2007). A study conducted on chronic lymphocytic and myeolocythic leukaemia patients, applied the lognormal distribution on chronic lymphocytic leukaemia diagnosed in Caucasian patients lived in Brooklyn between 1943 and 1952 (Feinleib and Macmahon, 1960).

### 2.5. Loglogistic Distribution

A random variable T has the Loglogistic distribution with the following hazard, density and survivorship functions

$$h(t\ ,\lambda,\gamma) = \frac{\lambda\,\gamma t^{\gamma-1}}{1 + \lambda\, t^{\gamma}}$$

$$f(t,\lambda,\gamma) = \frac{\lambda\,\gamma t^{\gamma-1}}{(1 + \lambda\, t^{\gamma})^2}$$

$$S(t,\lambda,\gamma) = \frac{1}{(1 + \lambda\, t^{\gamma})}$$

where λ>0, γ>0.

According to the study of Gupta et al., (1999) the loglogistic distribution is proved to be suitable in analyzing survival data conducted by Cox, Cox and Oakes, Bennet, O'Quinley and Sruthers. Gupta et al., used loglogistic distribution in survival analysis on lung cancer data in their studes. A study on the spreading ratio of HIV virus in San Fransisco between 1978 and 1986 indicated that loglogistic was the most suitable model among many distribution models to use with half censored data (Byers et al., 1988). A study emphasized that the maximum likelihood estimation was the most suitable method in estimating the parameters when performing analyses using loglogistic distribution on grouped data such as half-censored data (Zhou, Mi and Guo, 2007).

## 3. ANALYTICAL METHOD

### 3.1. Likelihood Ratio Test

A likelihood ratio test is a statistical test used to compare the fit of two models, one of which (the null models) is a special case of other (the alternative model). The test is on the likelihood ratio, which express how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model in favor of the alternative model. Both models are fitted to the data and their log-likelihoods are recorded.

The use of likelihood ratio in statistical inference is common (Edwards, 1972; Royall, 1997) and the role of likelihood in model comparison is well established (Akaike, 1973; Schwartz, 1978). Furthermore, the likelihood ratio plays a pivotal role in most approaches to hypothesis testing.

## 4. MODEL RESULTS

**Table 1. Parametric Regression model Fitted to Breast Cancer Data**

| S. No. | COVARIATES | EXPONENTIAL | | WEIBULL | | GOMPERTZ | | LOGNORMAL | | LOGLOGISTIC | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Haz. Ratio | S. E. | Haz. Ratio | S. E. | Haz. Ratio | S. E. | Coef. | S.E. | Coef. | S.E. |
| 1 | AGE | 1.00348 | 0.01211 | 1.00672 | 0.12158 | 1.00513 | 0.01211 | -0.00018 | 0.00763 | -0.00141 | 0.00751 |
| 2 | MENOPAUSE | 1.12916 | 0.28428 | 1.11200 | 0.28112 | 1.12882 | 0.28500 | -0.11880 | 0.15719 | -0.10835 | 0.15574 |
| 3 | HORMONE | 0.80688 | 0.13607 | 0.75180 | 0.12651 | 0.75167 | 0.12661 | 0.19069 | 0.10615 | 0.18698 | 0.10368 |
| 4 | SIZE | 1.01208* | 0.00484 | 1.01367* | 0.00489 | 1.01350* | 0.00489 | -0.00869* | 0.00331 | -0.00858* | 0.00312 |
| 5 | NODES | 1.04815* | 0.00998 | 1.05655* | 0.01015 | 1.05490* | 0.01015 | -0.03933* | 0.00825 | -0.38928* | 0.00797 |
| 6 | GRADE | 1.47335* | 0.21331 | 1.55249* | 0.22675 | 1.55477* | 0.22733 | -0.31943* | 0.09201 | -0.28392* | 0.90844 |
| 7 | PROG_RECP | 0.99498* | 0.00116 | 0.99461* | 0.00118 | 0.99469* | 0.00118 | 0.00287* | 0.00057 | 0.00307* | 0.00065 |
| 8 | ESTRG_RECP | 0.99984 | 0.00054 | 0.99748 | 0.00055 | 0.99981 | 0.00002 | 0.00019 | 0.00037 | 0.00013* | 0.00036 |
| | **DEVIANCE** | 866.62246 | | 812.7804 | | 834.75706 | | 798.36414 | | 804.0081 | |

**Table 2. Difference in deviance of various models compared to Lognormal**

| S. NO. | DISTRIBUTIONS | LR |
|---|---|---|
| 1 | EXPONENTIAL | 68.25832 |
| 2 | WEIBULL | 14.41626 |
| 3 | GOMPERTZ | 36.39292 |
| 4 | LOGLOGISTIC | 5.64396 |

## 5. RESULTS

The parametric models were fitted using STATA 12 and the results are presented in Table 1. From the table we see that the four covariates namely size, nodes, grade and Progesterone Receptors are significantly associated with the survival time under all the model assumptions. Among the models, the lognormal has the lowest level of deviance compared to all other models. The difference in deviance of the all other models compared to lognormal are given in Table 2. It is further noted that all other models have significantly higher deviance compared to lognormal.

## 6. DISCUSSION AND CONCLUSION

A study conducted by Nakhaee and Law (2011), showed that, survival following a diagnosis of HIV infection was modeled by applying parametric survival models on people who were only diagnosed with HIV or HIV and AIDS registered in the national surveillance system from 1997 to 2003. likelihood based criteria for model selection indicated that the Weibull model was the best fitting parametric model for predicting survival following both HIV and AIDS diagnoses. Baghestani et al., (2010) the results indicated that the early detection of a cancer at a young patient age and in primary stages is important to increase survival from gastric cancer. According to statistical criteria, a parametric model can also be useful statistical model to find prognostic factors in the presence of interval censoring. Deviance supported the loglogistic regression as the best option.

Pourhoseingholi et al., (2007) pointed out, in multivariate models, Cox and Exponential are the same with respect to AIC and standardized variability. But in univariate, all parametric ones are better than Cox except for tumor size and the lognormal is the first choice among parametric models. Jiezhi Qi (2009), viewed that, after comparison of all the models and assessment of goodness of fit, found that the loglogistic AFT model fits better, for randomized placebo-controlled trial to prevent Tuberculosis in Uganda adults infected with HIV.

Ponnuraja and Venkatesan (2010), applied likelihood based criteria for model selection indicated that the Gamma model was the best fitting parametric model for tuberculosis clinical trial data. Hayat et al., (2010), showed that age was not found as a risk factor. By using AIC, Gompertz model was more suitable, for Breast Cancer Registry Data from Ege University Cancer Research Center.

Inspite of different models suggested and recommended by many researchers, who followed different methodologies in their experiments, the result of our study has a strong inclination for the Lognormal method as the most suitable one; better than other based on Deviance. Further studies are needed to confirm the findings.

**Acknowledgments**

I wish to express my gratitude and thanks to my guide and my parents, family members for their valuable support and cooperation.

## REFERENCES

[1]. Ahuja J. C. and Nash S. W. (1967), "The generalized Gompertz verhulst family of distributions", Sankhya ser A, 29(2), 141-56.

[2]. Ahmad Reza Baghestani, Ebrahim Hagizadeh, and Seyed Reza Fatemi (2010), "Parametric model to analyse the survival of gastric cancer in the presence of interval censoring", Tumor, 69, 433-7.

[3]. Akturk Hayat, Asli Suner, Burak Uyar, et al (2010), "Comparison of Five Survival Models: Breast Cancer Registry Data from Ege University Cancer Research Center", Turkiy, J Med Sci , 30(5), 1665-74.

[4]. Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle". In B. N. Petrov & F. Csaki (Eds.), second international symposium on information theory, 267-81. Budapest: Academia kiao.

[5]. Byers, R. H., Morgan, W. M., Darrow, et al (1988), "Estimating AIDS infection rates in the San Francisco cohort", AIDS , 2(3), 207-10.

[6]. Brown, K. and Forbes, W. (1974), "A mathematical model of aging processes", Journal of Gerontology , 29 (1), 46–51.

[7]. Benjamin, Bernard and Haycocks et al (1980), "The Analysis of Mortality and Other Actuarial Statistics" , London: Heinemann.

[8]. Bemmaor, Albert C.; Glady, Nicolas (2012), "Modeling Purchasing Behavior With Sudden 'Death': A Flexible Customer Lifetime Model" Management Science, 58 (5), 1012–21.

[9]. Cox D. R. (1972), "Regression model and life tables" (with discussion), J. Royal Stat. Soc. (B), 34, 187-220.

[10]. Cox, D. R. and Oakes, D (1984,. "Analysis of Survival Data", Chapman and Hall.

[11]. Dumonceaux, R. and C. E. Antle and Hass G. (1973), "Likelihood Ratio Test for Discrimination between Two Models With Unknown Location and Scale Parameters", Technometirc, 15, 19-27.

[12]. Elisa, T. Lee and John Wenyu Wang (2003), "Statistical Methods for Survival Data Analysis", Third Edition, John Wiley and Sons Inc., Newyork.

[13]. Edwards, A (1972), "Likelihood", London Cambridge University Press.

[14]. Economos, A (1982), "Rate of aging, rate of dying and the mechanism of mortality", Archives of Gerontology and Geriatrics, 1 (1), 46–51.

[15]. Feinleib M and Macmahon B (1960), "Variation in the duration of survival of patients with the chronic leukemias", Blood 1960, 15(3), 332-49.

[16]. Gupta R.C., Akman O and Lvin S (1999), "A study of log-logistic model in survival analysis", Biom J, 41(4), 431-43.

[17]. Gore, S. M., Pocock, S. J., and Kerr, G. R. (1984), "Regression models and nonproportional hazards in the analaysis of breast cancer survival", Appl. Stat., 33, 176-95.

[18]. Horner R.D (1987), "Age at onset on Alzheimer's disease: clue to the relative importance of etiologic factors?", Am J Epidemio, 126(3), 409-14.

[19]. Hosmer, D. W. and Lemeshow, S. and May, S. (2008), "Applied Survival analysis: Regression Modeling of Time to Event Data", Second Edition, John Wiley and Sons Inc., Newyork.

[20]. Inghelmann R, Grande E, Francisci S et al (2005), "National estimates of cancer patients survival in italy: A model-based method", Tumori, 91(2), 109-15.

[21]. Jin, Z., Lin, D. Y., Wei, L. J et al (2003)., "Rank-based inference for the accelerated failure time model", Biometrika, 90, 341-53.

[22]. Jiezhi Qi (2009), "Comparison of Proportional Hazards and Accelerated Failure Time models", Thesis, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

[23]. Kay, R and Kinnersley, N. (2002), "On the use of the accelerated failure time model as an alternative to the propotional hazard model in the treatment of time to event data : A case study in Influenza", Drug information Journal, 36, 571-79.

[24]. Kalbfleisch, J. D. and Prentice, R. L. (1980), "The Statistical Analysis Failure Time Data", Wiley, New York.

[25]. Kalbfleisch, J. D. and Prentice R. L. (1980) "The Statistical analysis of failure time data", John Wiley and Sons Inc., Newyork.

[26]. Kleinbaum, D. G. (1996), "Survival analysis: A self learning text", Springer-Verlag, New York.

[27]. Limpert E, Stahel W.A., Abbt M. (2001), "Lognormal distributions across the sciences: Keys and clues" BioScience, 51(5), 341-52.

[28]. Mann, N. R. Schafer, R. E. and Singpurwalla, N. D. (1974), "Methods for Statistical analysis of reliability and life data", John Wiley and Sons, New york.

[29]. Mohamad Amin Pourhoseing holi, Ebrahim Hajizadeh, Bijan Moghimi Dehkordi. et al (2007),"Comparing Cox Regression and Parametric models for survival of patients with Gastric Carcinoma", Asian pacific J Cancer Prev, 8 ,412-16.

[30]. Nakhaee, F. and Law, M. (2011,. "Parametric modeling of survival following HIV and AIDS in the era of highly active anti retroviral therapy: data from Australia", Eastern Mediterranean Health Journal, 7(3), 2011.

[31]. Nelson W. (1982), "Applied Life Data Analysis, John Wiley and Sons", New York.

[32]. Ohishi, K., Okamura, H. and Dohi, T. (2009), "Gompertz software reliability model: estimation".

[33]. Prathap et al (2011), "Compariaon of Reliability Models for Life time Data", Recent Trends in Statistics and Computer Application, Manonmaniam Sundaranar University, pp. 163-170 ISBN 978-93-81402-12-2.

[34]. Ponnuraja, C. and Venkatesan, P. (2010), "Survival models for exploring tuberculosis clinical trial data –an empirical comparison", Indian Journal of Science and Technology, 2(7), 755-58.

[35]. Preston, Samuel H.; Heuveline, Patrick and Guillot, Michel (2001), "Demography:measuring and modeling population processes",Oxford: Blackwell.

[36]. Pugh, E. L. (1963), "The Best Estimate of Reliability in the Exponential case", Journal of the Operation Research Society of America, 11, 56-61.

[37]. Royall, R. M. (1997), "Statistical evidence: A likelihood paradigm", London. Chapman & Hall.

[38]. Schwartz, G. (1978), "Estimating the dimension of a model. Annals of statistics", 6, 461-64.

[39]. Tai P, Chapman J.A., Yu E, Jones D, Yu C, Yuvan F, et al.(2007), "Disease-specific survival for limited-stage small –cell lung cancer affected by statistical method of assessment", BMC cancer, 7:31

[40]. Viscomi S, Pastore G, Dama E, et al. (2006), "Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukemia", Ann Oncol, 17(1), 167-71.

[41]. Vaupel, James W. (1986), "How change in age-specific mortality affects life expectancy" Population Studies, 40 (1), 147–57.

[42]. Willemse, W. J and Koppelaar, H. (2000), "Knowledge elicitation of Gompertz' law of mortality", Scandinavian Actuarial Journal, 2, 168–79.

[43]. Weibull W. (1951), "A statistical distribution functions of wide applicability", J Appl Mech, 18(2), 293-97.

[44]. Zhou,Y.Y., Mi, J. and Guo, S. (2007), "Estimation of parameters in logistic and log-logistic distribution with grouped data. Life time data",Anal., 13(3), 421-29.

[45]. Zelen, Marvin and Mary C. Dannemiller (1961), "The Robustness of Life Testing Procedures Derived from the Exponential Distribution", Technometrics, 3, 29-49.