Bayesian survival analysis: Comparison of survival probability of hormone receptor status for breast cancer data

Article *in* International Journal of Data Analysis Techniques and Strategies · January 2017

DOI: 10.1504/IJDATS.2017.083061

Esin Avci
Giresun University
27 PUBLICATIONS **17** CITATIONS

Bayesian survival analysis: comparison of survival probability of hormone receptor status for breast cancer data

Esin Avcı

Department of Statistics,
University of Giresun,
Giresun, 28000, Turkey
Email: esinavci@hotmail.com

Abstract: Survival analysis is a family of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. The Cox model is the most widely used survival model in health sciences, but it is not the only model, parametric models in which the distribution of the event is specified in terms of unknown parameters. Over the last few years, there has been increased interest shown in the application of survival analysis based on Bayesian methodology. In this article, we consider Bayesian survival analysis to compare survival probability of hormone receptor status for breast cancer based on lognormal distribution estimated survival function. The Bayesian approach is implemented using WinBugs.

Keywords: Bayesian survival analysis; survival function; hormone receptor status; breast cancer.

Reference to this paper should be made as follows: Avcı, E. (2017) 'Bayesian survival analysis: comparison of survival probability of hormone receptor status for breast cancer data', *Int. J. Data Analysis Techniques and Strategies*, Vol. 9, No. 1, pp.63–74.

Biographical notes: Esin Avcı is an Assistant Professor specialising in Statistics. She held previous faculty positions at Sinop University as a Research Assistant. Her areas of interest and expertise are in Bayesian survival analysis; and in generalised statistical methods involving count data that contain data dispersion.

1 Introduction

Survival analysis deals with analysis of time duration to until one or more events happen. The Cox proportional hazards (PHs) model is popular model for analysing survival data because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. The utility of this model stems from the fact that few assumption are needed to determine hazard ratios based on the coefficients. The coefficient is easily interpreted and clinically meaningful (Hosmer and Lemesow, 1989). There are parametric survival models for which the restrictive assumption of hazards is not required. A parametric survival model is one in which survival time is assumed to follow a known distribution. Exponential, Weibull, lognormal, log-logistic and generalised gamma distributions are most commonly used. Many parametric models are acceleration failure time (AFT) models rather than PH models. The interpretation of parameters differs for AFT and PH models. The AFT assumption is applicable for a comparison of survival times whereas the PH assumption is applicable for a comparison of hazards (Kleinbaum and Klein, 1996). Parametric models have some advantages over Cox models in general. With small sample size, relative efficiencies may further change in favour of parametric models (Nardi and Schemper, 2003). Over the last few years, there has been increased interest shown in the application of survival analysis based on Bayesian methodology. Researchers did not use Bayesian analysis frequently in medical studies because it has a complex theory. Bayesian analysis of survival data has received much recent attention due to advances in computational and modelling techniques (Ibrahim et al., 2001). Bayesian survival analysis consists of data and prior information. It generates conclusion based on the synthesis of new information from an observed data and historical knowledge or expert opinion. Historical knowledge from past similar studies can be very helpful in interpreting the results of the current study. Therefore, Bayesian survival analysis reflects researches subjective beliefs. Prior elicitation plays the most crucial role in Bayesian survival analysis. Bayesian survival analysis cannot be used for any modelling without using a prior distribution (Ibrahim et al., 2001; SAS Institute, 2006). Parametric models play an important role in Bayesian survival analysis, since many Bayesian analysis in practice are carried out using parametric models (Ibrahim et al., 2001).

Recently, few works dealing with application to medicine or public health published on the Bayesian survival analysis methods. Khan and Khan (2013) analysed survival time of patient treated with two different treatment using Weibull parametric model of Bayesian survival analysis. Omurlu et al. (2009) compare performance of Cox regression analysis and Bayesian survival analysis under varying sample sizes using Monte Carlo (MC) simulation and for disease-free survival in breast cancer patients. Yin and Ibrahim (2006) analysed a simulation study using Bayesian survival analysis for varying sample sizes. Wong et al. (2005) used Bayesian to investigate the effectiveness of silver diamine fluoride and sodium fluoride varnish in arresting active dentine caries in Chinese pre-school children.

In this study, I first introduce briefly in Section 2 survival analysis then commonly used parametric survival models (lognormal, Weibull, exponential, and log-logistic) and finally, Bayesian approach for survival analysis. Section 3 gives some results and survival curves of breast cancer data obtained via WinBugs and Section 4 contain discussion and conclusions.

2 Bayesian survival analysis

2.1 Survival analysis

Let T be a continuous non-negative random variable representing the survival times of individuals in some population. Let $f(t)$ denote the probability density function (pdf) of T and let the distribution function be

$$F(t) = P(T \leq t) = \int_0^t f(u)du \tag{1}$$

The probability of an individual's surviving till time t is given by the survivor function:

$$S(t) = 1 - F(t) = P(T > t) \tag{2}$$

Survival data are often right censored. This means that, survival times are known for only a portion of the individuals under study, and the reminder of the survival times are known only to exceed certain values. The likelihood function for right censored data

$$L = \prod\nolimits_{i=1}^{n} \left[f(t_i) \right]^{\delta_i} \left[S(t_i) \right]^{1-\delta_i} \tag{3}$$

where δ_i is an indicator variable which takes value 1 if observation is failed and 0 if censored.

There are two main models in survival analysis; Cox PH (semiparametric model) and parametric model. The Cox PH model has become a standard tool for constructing multiple regression models for assessing the influence of prognostic factors in censored survival data. The dominance of the Cox model depends on two factors. First, for implementing Cox model rapid improvements in a sowftware appeared. The second is the convenience and robustness of not needing to specify the survival distribution (Royston, 2001). Although obviously useful, the Cox model has drawbacks. The PHs assumption may fail, particularly in datasets with long follow-up times (Valsecchi et al., 1996). There are parametric survival models for which the restrictive assumption of PHs is not required. Parametric survival models could provide a more suitable description of the data if one is able to identify the distribution of the survival time (Khanal et al., 2014). A further point is that estimates from a Cox model are inefficient compare with those from a correctly specified parametric model (Royston, 2001). Some parametric models are accelerated failure time (AFT) models which assume that the relationship between the logarithm of survival time and covariates is linear (Lee and Wang, 2003). Most commonly used exponential, Weibull, log-logistic, lognormal and generalised gamma models. Exponential and Weibull parametric models can work both in PH and AFT metric. Log-logistic, lognormal and generalised gamma models work only in AFT metric.

The methods of assessing the adequacy of data to a specific probability distribution or of selecting the best parametric survival model among several competitors can generically be grouped into plotting procedures and formal methods for goodness of fit evaluation. Plotting procedures are common methods used to explore and visually inspect the adequacy of a specific survival time distribution. In particular, use a probability plot to assess whether a particular distribution fits data. To determine which distribution best fits data, comparing how closely the plot points lie to the best-fit lines of a probability plot. Another fundamental indicators to identify appropriate parametric survival model is based on the shape of the baseline hazard function.

There are several well-known formal goodness of fit tests and tests for discriminating among models when checking the adequacy of parametric survival models. In this study, we formally measured model adjustment using deviance and the Akaike's information criteria (AIC). These measures enable comparisons between not nested models. Smaller values of both measures indicate better adjusted models.

2.2 Lognormal model

One of the commonly used parametric survival model is the lognormal model. For this model, we assume that the logarithms of the survival times are normally distributed. If t has a lognormal distribution with parameters (μ, σ^2), denoted by $LN(\mu, \sigma^2)$, then

$$f\left(t \mid \mu, \sigma^2\right) = (2\pi)^{-\frac{1}{2}}(t\sigma)^{-1}\exp\left\{-\frac{1}{2}\left(\log(t)-\mu\right)^2\right\} \tag{4}$$

The survival function is given by

$$S(t \mid \mu, \sigma) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right) \tag{5}$$

$\Phi(.)$ is the standard normal distribution function. The parameter μ is the log geometric mean. The skewness and kurtosis of t are monotonic function of σ.

The hazard function increases initially to a maximum and then decreases (almost as soon as the median is passed) to zero as time approaches infinity (Watson and Wells, 1961). Therefore, the lognormal distribution is suitable for survival patterns with initially increasing and then decreasing hazard rate (Lee and Wang, 2003).

Typically, μ is modelled as a linear function of covariates $\boldsymbol{x}$, so that

$$lnt = \beta_0 + \beta^T \boldsymbol{x} + e\sigma \tag{6}$$

where $e \sim N(0, 1)$. Estimation of the parameters β_0 and β from a sample of n_U uncensored and $n - n_U$ censored observation is done by maximising the log likelihood function, which is

$$lnL = \sum_{i=1}^{n_U} lnf(t) = \sum_{i=n_U+1}^{n} lnS(t) \tag{7}$$

When β is small, it can be interpreted as the percentage increase if $\beta > 0$ or percentage decrease if $\beta < 0$ in the average survival time and/or median survival time when we increase the covariate value of $\boldsymbol{x}$ by one unit (Zhang, 2005).

2.3 Exponential model

The exponential distribution is characterised by a constant hazard rate λ, its only parameter. A high λ value indicates high risk and short survival; a low λ value indicates low risk and log survival. When the survival time t follows the exponential distribution with a parameter λ, the pdf is defined as;

$$f(t) = \lambda e^{-\lambda t} \tag{8}$$

and the survival function is then,

$$S(t) = e^{-\lambda t} \tag{9}$$

So that, the hazard function is

$$h(t) = \lambda \tag{10}$$

a constant, independent of t (Lee and Wang, 2003).

2.4 Weibull model

The Weibull distribution is a generalisation of the exponential distribution. However, unlike the exponential distribution, it does not assume a constant hazard rate and therefore has broader application.

The Weibull distribution is characterised by two parameters, γ and λ. The value of γ determines the shape of the distribution curve and the value of λ determines its scaling. Consequently, γ and λ are called the shape and scale parameters, respectively (Lee and Wang, 2003).

The pdf is,

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}} \tag{11}$$

the survival function is then,

$$S(t) = e^{-(\lambda t)^{\gamma}} \tag{12}$$

the hazard function is,

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \tag{13}$$

2.5 Log-logistic model

The survival time *t* has a log-logistic distribution if log(*t*) has a logistic distribution. The density, survival and hazard functions of the log-logistic distribution are, respectively (Lee and Wang, 2003),

$$f(t) = \frac{\alpha\gamma t^{\gamma-1}}{\left(1+\alpha t^{\gamma}\right)^{2}} \tag{14}$$

$$S(t) = \frac{1}{1+\alpha t^{\gamma}} \tag{15}$$

$$h(t) = \frac{\alpha\gamma t^{\gamma-1}}{1+\alpha t^{\gamma}} \tag{16}$$

2.6 Bayesian approach

Bayesian analysis generates conclusions based on the synthesis of new information from the observed data and previous knowledge or external evidence (Wong et al., 2005).

In Bayesian approach, a probability model for observed data *y*, given a vector of unknown parameters β, leading to the likelihood function $L(y \mid \beta)$. Then, we assume that β is random and has a prior distribution denoted by $p(\beta)$. Inference concerning β is then based on the posterior distribution of β is proportional to the likelihood multiplied by the prior is given by

$$p(\beta \mid y) \propto L(y \mid \beta)p(\beta) \tag{17}$$

and thus it involves a contribution from the observed data through $L(y \mid \beta)$, and a contribution from prior information quantified through $p(\beta)$ (Ibrahim et al., 2001).

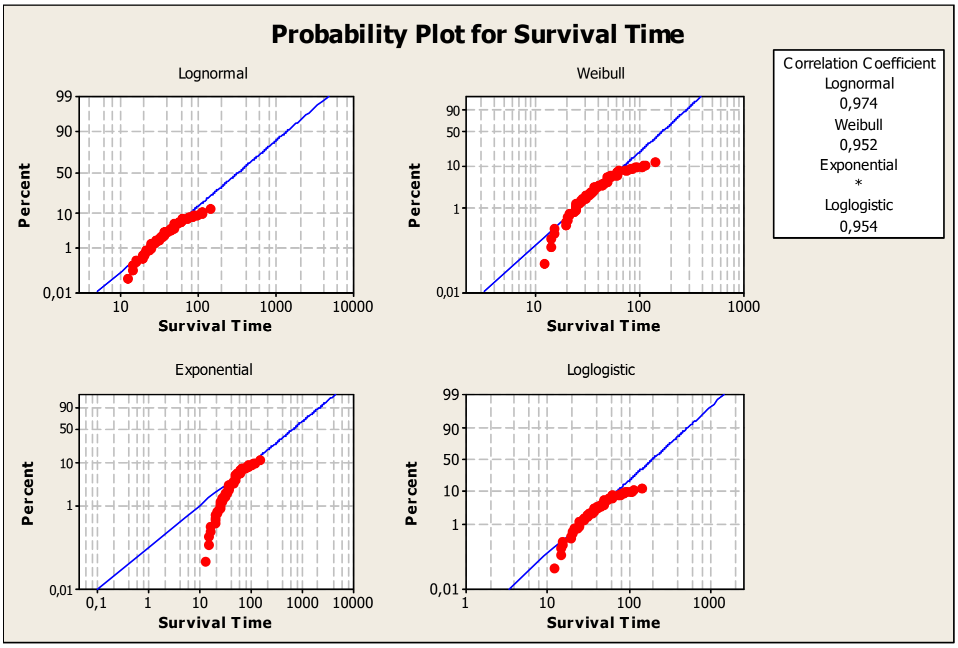
3 Analysis of breast cancer data

A retrospective analysis is performed to 1,464 women diagnosed with breast cancer in Istanbul between 1993 to 2002. The data is taken from original performance of hospital admitted breast cancer patients from the breast surgery clinic of the Istanbul university medical faculty. Survival time defined as a period between the diagnosis of disease and death or the end of patient's follow-up. Age, estrogen and progesterone receptor status are investigated as prognostic factors that associated with breast cancer survival time. Summary for the entire patient population is shown in Table 1.

Table 1 Summary of patients

Variable		*Mean ± Stdv.*	
Age		50.95 ± 12,74	
Variable	*Statu*	*Total*	*Died (%)*
Overall		1,464	93 (0.06)
Estrogen receptor status (ER)	Negative	555	66 (0.12)
	Positive	909	27 (0.03)
Progesteron receptor status (ER)	Negative	762	74 (0.10)
	Positive	702	19 (0.03)

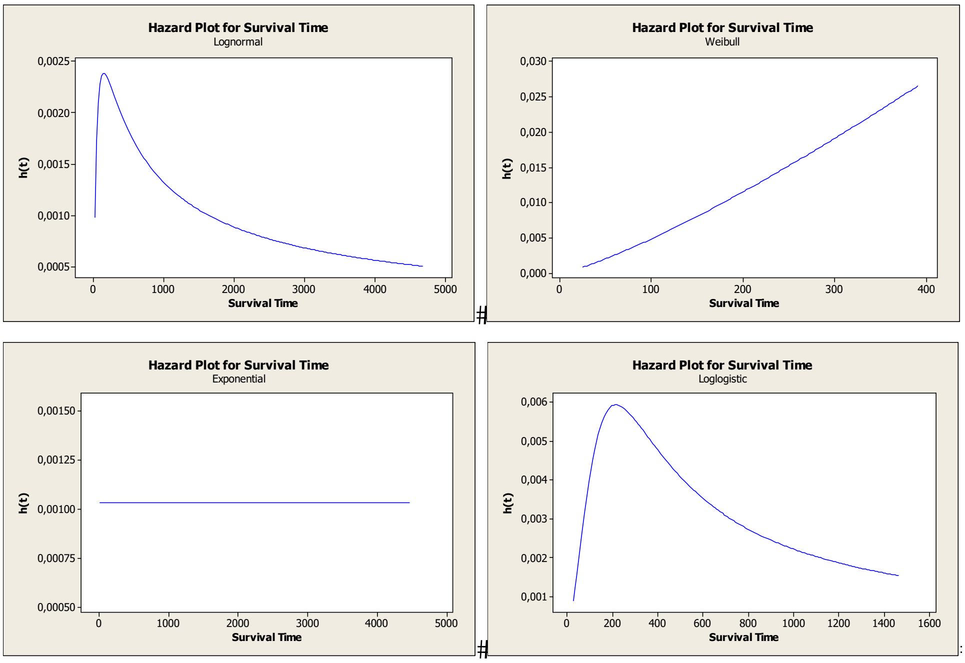
Figure 1 Probability plots for breast cancer data (see online version for colours)



To identify the appropriate parametric model, probability plot for commonly used parametric survival models (lognormal, Weibull, exponential, and log-logistic) plotted by using Minitab's distribution ID plot (right censoring) option (Figure 1). In order to make conclusion that which survival distribution is best for cancer data, four common survival distributions are compared according to their how closely the plot points lie to the best-fit lines of a probability plot from Figure 1, the best distribution fits for breast cancer data is lognormal distribution (the biggest correlation coefficient).

Another exploratory technique is through plot of hazard plot for four distributions is given in Figure 2.

Figure 2 Hazard plots for breast cancer data (see online version for colours)



As further check cumulative hazard plots which are used for visually examining distributional model assumption is given in Figure 2. Figure 2 shows that the hazard function increases initially to a maximum and then decreases to zero as time approaches infinity.

Table 2 presents the deviances and AIC for four well-known parametric models fitted to the breast cancer data.

Table 2 Deviances and AIC for four parametric models fitted to the breast cancer data

Distribution	*Deviance*	*AIC*
Lognormal	1,877.42	1,472.02
Weibull	1,887.04	1,539.86
Exponential	1,977.27	1,488.62
Log-logistic	1,897.84	1,555.54

The appropriateness of the lognormal distribution for breast cancer data is examined statistically with deviance and AIC and graphically with probability and hazard plots.

Depending on aspects there is a benefit (relative efficiencies) of parametric survival models over Cox regression, Cox model is not used. To assessment of influence of age, estrogen and progesterone receptor status on survival time of breast cancer, lognormal survival analysis based on Bayesian approach is used. The Cox basic infrastructure for the analysis is done by using the commercially available WinBugs software employing the Markov chain Monte Carlo (MCMC) methodology. The Bugs language allows a concise expression of the parametric model to denote stochastic relationships and to denote deterministic relationships. Thus, using a MCMC method of parameter estimation with non-informative priors, one is able to obtain the posterior estimates and credible regions of estimates of these effects. Graphical displays of convergence history and posterior densities affirm the stability of the results.

Table 3 presents posterior estimation and credible regions with normal priors. To determine number of iterations to obtain samples that can be used for posterior inference. By saving the more samples, the more accurate will be obtained in posterior estimation. One way to assess the accuracy of the posterior estimates is by calculating the MC error for each parameters. As a rule of thumb, the simulation should be run until the MC error for each parameter of interest is less than about 5% of the sample standard deviation (SD). Iterating process is carried beyond 25,000 that the estimates proved to be very stable. To reduce the potential bias, the first 2,500 samples as burn-in is discarded. All the parameters 95% credible regions does not contain zero. So all the parameter had significant effect on survival time of breast cancer. Patients with positive estrogen and progesterone status, have better survival (the patients' average survival time will increase by about 44% and 57%, respectively).

Table 3 Posterior estimates and credible regions

Parameter	*Mean*	*SD*	*MC error*	*2.50%*	*97.50%*	*Exp (β)*
Age	0.0658	0.0009	0.00001	0.06395	0.0675	1.0680
ER	0.3642	0.0723	0.0012	0.224	0.5078	1.4394
PR	0.4489	0.0691	0.0009	0.3133	0.5844	1.5666

Figure 3 displays posterior densities and convergence history for all three parameters. Note that for all the parameters have symmetry in the posterior densities. The history plots looks like nice oscillograms around a horizontal line without any trend. The Markov chain is most likely to be sampling from the stationary distribution and is mixing well.

In order to make conclusion that which receptor status is better for patients of breast cancer, survival curves based on Bayes estimate of lognormal survival function was used. It could be seen from Figure 4 that the patients receiving esterogen and progesterone receptor had higher survival probabilities than those not receiving.

Figure 3 Marginal posterior densities and history plots of parameters (see online version for colours)

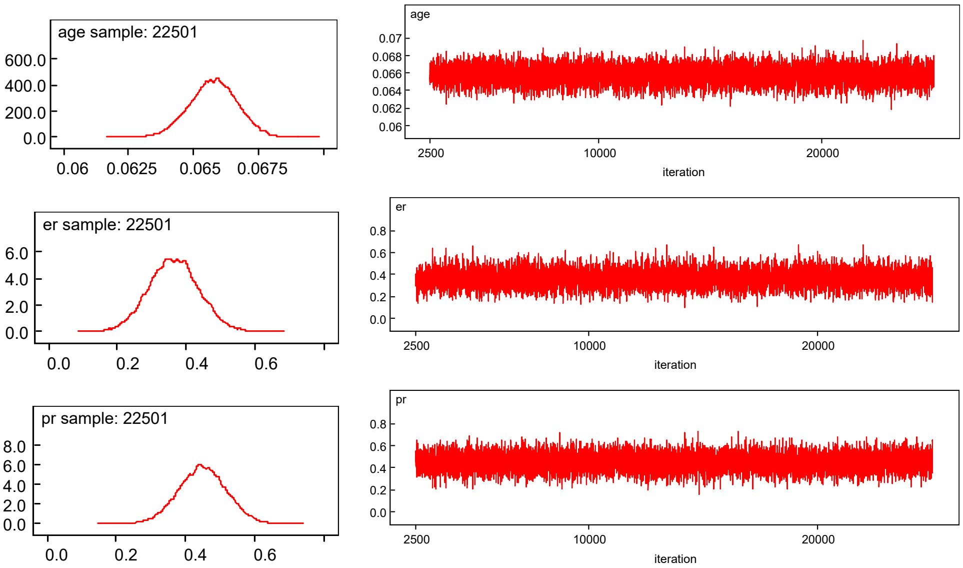


Figure 4 Estimated survival curve of estrogen and progesterone

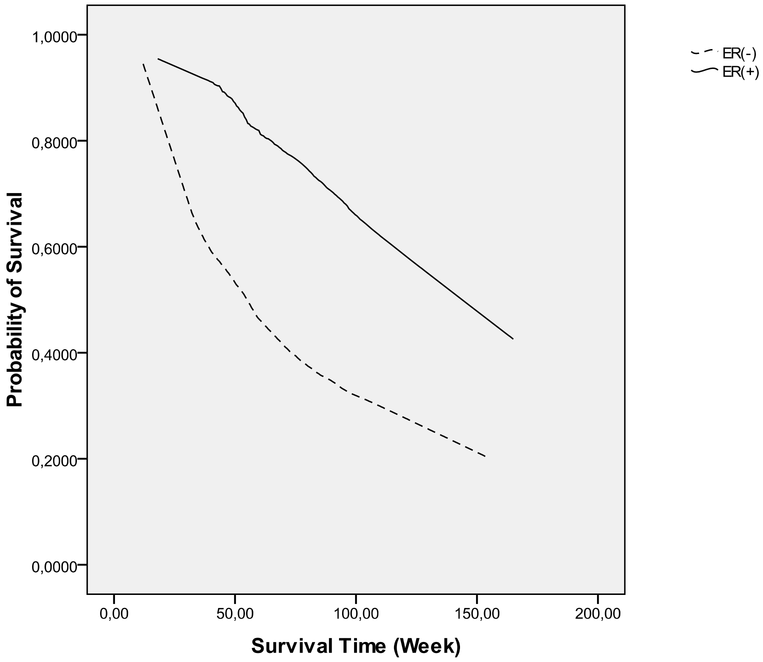
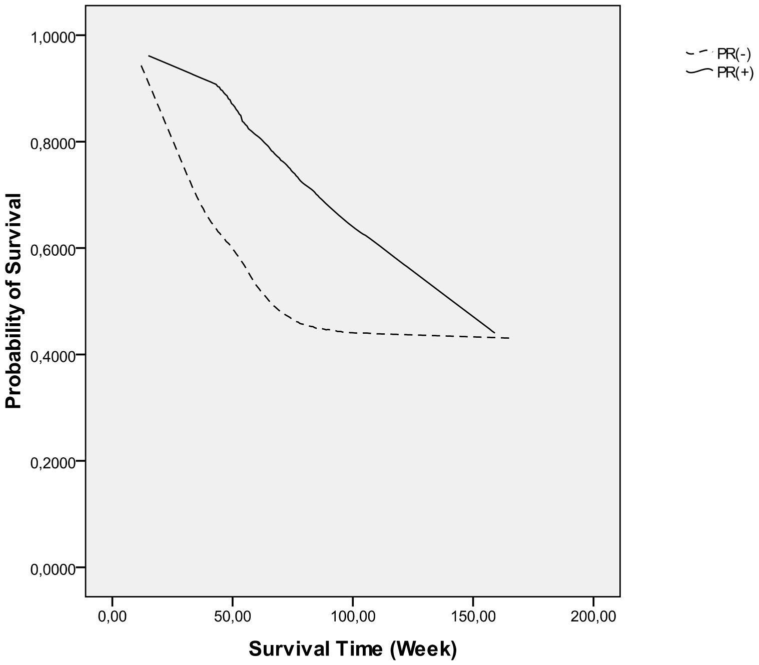


Figure 4 Estimated survival curve of estrogen and progesterone (continued)



4 Discussion and conclusions

In the area of medical research, the widely used regression model for time to event data is Cox PH model because of its familiarity and convenience (Anderson, 1991; Cox, 1972). Parametric models are attractive option when either hazard function themselves are of primary interest, or when relative times instead of relative hazards are the relevant measures of association (Cox et al., 2007). Ease of interpretation of parameters may be another benefit especially for clinicians. However, the estimation of these models is carried out by assuming a distribution of the survival time. If the distribution of survival time is not recognised, then estimation based on parametric models becomes questionable.

There are few works about lognormal survival model. Royston (2001) compared Cox and lognormal model for breast and ovarian cancer data. John et al. (1994) compared three of lognormal models with log-rank test. Chapman et al. (2008) used lognormal survival analysis for breast cancer patients who received adjuvant endocrine therapy. Tai et al. (2004) used lognormal survival model to estimate survival of metastatic breast cancer. Khanal et al. (2014) used lognormal survival model to estimate survival of acute liver patients. Works dealing with application to medicine or public health published on the Bayesian survival analysis methods are as follows. Khan and Khan (2013) analysed survival time of patient treated with two different treatment using Weibull parametric model of Bayesian survival analysis. Omurlu et al. (2009) compared performance of Cox regression analysis and Bayesian survival analysis under varying sample sizes using MC simulation and for disease-free survival in breast cancer patients. Yin and Ibrahim (2006) analysed a simulation study using Bayesian survival analysis for varying sample sizes. Wong et al. (2005) used Bayesian to investigate the effectiveness of silver diamine fluoride and sodium fluoride varnish in arresting active dentine caries in Chinese pre-school children. Khan et al. (2005) used the Bayesian framework to derive the predictive distributions of future responses given both Type II censored samples and Type II median censored samples, assuming non-informative and informative prior distributions for the parameters.

This article applied lognormal survival analysis based on Bayesian approach to breast cancer data. To determine breast cancer data distribution deviance and AIC as test statistics and probability and hazard plot as graphical methods are used. Statistically and graphically distribution of the data clearly matching with lognormal distribution. Age, estrogen and progesterone receptor status handled as prognostic factors. All these prognostic factors are found significant effect on breast cancer patients'. Being provided with visuals of history plot and density structure of the parameters of interests allows one to logically determine if the convergence is following a logical pattern and not deviating wildly as one goes through the iterations. Receiving estrogen and progesterone receptor increases probability of survival about 44% and 57%, respectively. Survival curves are drawn from survival probabilities which are obtained from Bayesian approach of lognormal survival model. As shown from survival curve receiving estrogen and progesterone receptor had higher survival probabilities than those not receiving.

References

Anderson, P.K. (1991) 'Survival analysis 1982–1991: the second decade of proportional hazards regression model', *Statist. Med.*, Vol. 10, No. 12, pp.1931–1941.

Chapman, J.W., Meng, D., Shepherd, L., Parulekar, W., Ingle, J.N., Muss, H.B., Palmer, M., Yu, C. and Goss, P.E. (2008) 'Competing causes of death from a randomized trail of extended adjuvant endocrine therapy for breast cancer', *JNCI*, Vol. 100, No. 4, pp.252–260.

Cox, C., Chu, H., Schneider, M.F. and Munoz, A. (2007) 'Parametric survival analysis and taxonomy of hazard functions for the genealized gamma distribution', *Statist. Med.*, Vol. 26, No. 23, pp.4352–4374.

Cox, D.R. (1972) 'Regression models and life tables (with discussions)', *Journal of the Royal Statistics Society Series B*, Vol. 34, No. 2, pp.187–220.

Hosmer, D. and Lemesow, S. (1989) *Applied Survival Analysis*, Wiley, New York.

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001) *Bayesian Survival Analysis*, Springer-Verlag, New York.

John, W., Gamel, M.D., Robert, L.V., Pinuccia, V. and Bonadonna, G. (1994) 'Parametric survival analysis of adjuvant therapy for stage II breast cancer', *Veterans Administration Medical Center Cancer*, Vol. 74, No. 9, pp.2483–2490.

Khan, H.M.R., Haq, M.S. and Provost, S.B. (2005) *Bayesian Prediction for the Log-normal Model Under Type II Cencoring*, CAMS Report 0405-17, Spring.

Khan, Y. and Khan, A.A. (2013) 'Bayesian survival analysis of regression model using Weibull', *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 2, No. 12, pp.7199–7204.

Khanal, S.P., Sreenivas, V. and Acharya, S.K. (2014) 'Accelerated failure time models: an application in the survival of acute liver failure patients in India', *International Journal of Science and Research*, Vol. 3, No. 6, pp.161–166.

Kleinbaum, D.G. and Klein, M. (1996) *Survival Analysis: A Self-Learning Text*, Springer, USA.

Lee, E. and Wang, J. (2003) *Statistical Methods for Survival Data Analysis*, Wiley, New Jersey.

Nardi, A. and Schemper, M. (2003) 'Comparing Cox and parametric models in clinical studies', *Stat. Med*., Vol. 22, No. 23, pp.3597–3610.

Omurlu, I.K., Ozdamar, K. and Ture, M. (2009) 'Comparision of bayesian survival analysis and Cox regression analysis in simulated an breast cancer data sets', *Expert Systems with Applications*, Vol. 36, No. 8, pp.11341–11346.

Royston, P. (2001) 'The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors', *Statistica Neerlandica*, Vol. 55, No. 1, pp.89–104.

SAS Institute (2006) *Preliminary Capabilities for Bayesian Analysis in SAS/STATR Software*, SAS Institute Inc., Cary, NC, USA.

Tai, P., Yu, E., Vinh-Hung, V., Gserni, G. and Vlastos, G. (2004) 'Survival of patients with metastatic breast cancer: twenty-year data from two SEER registries', *BMC Cancer*, Vol. 4, No. 60, pp.1471–2407.

Valsecchi, M.G., Silvestri, D. and Sasieni, P.D. (1996) 'Evaluation of long term survival: use of diagnostics and robust estimators with Cox's proportional hazards models', *Statistics in Medicine*, Vol. 15, No. 24, pp.2763–2780.

Watson, G.S. and Wells, W.T. (1961) 'On the possibility of improving the mean useful life of items by eliminating those with short lives', *Technometrics*, Vol. 3, No. 2, pp.281–298.

Wong, M.C.M., Lam, K.F. and Lo, E.C.M. (2005) 'Bayesian analysis of clustered interval-cencored data', *Journal of Dental Research*, Vol. 84, No. 9, pp.817–821.

Yin, G. and Ibrahim, J.G. (2006) 'Bayesian transformation hazard model', in *Proceedings of the Second Lehmann Symposium-Optimality, IMS Lecture Notes-Monograph Series 2006*, Vol. 49, pp.170–182.

Zhang, D. (2005) *Analysis of Survival Data (ST745)*, Spring [online] http://www4.stat.ncsu.edu/~dzhang2/st745/syllabus.pdf.