Journal of Modern Mathematics and Statistics 4 (4): 127-136, 2010
ISSN: 1994-5388

A Model for Measuring Association Between Bivariate Censored Outcomes

A.F. Fagbamigbe and A.S. Adebowale
Department of Epidemiology, Medical Statistics and Environmental Health,
Faculty of Public Health, College of Medicine,
University of Ibadan, Nigeria

Abstract: The dependence between two random variables is completely described by their bivariate distribution. Bivariate survival analysis arises in the time to events analysis of measurements that are paired. Although, there are several consistent estimators of the bivariate distribution function, an efficient and consistent estimation has proven to be a difficult problem. It is of interest to determine if it exists, the possible association between pairs of variables, both of which are subject to censoring with recurrence times of kidney infection as a case study. Copula models which is one of the existing methods of measuring the possible association between bivariate censored variables were reviewed. The overall average recurrence time and its standard deviation are 102 and 131, respectively though the recurrence time in the first kidney has average and standard deviation of 112 and 144.01, respectively while the average and standard deviation of recurrence time in the second kidney recurrence time is 92 and 117.20, respectively. The study also showed that the modal recurrence time in the 2 kidneys is 42. The correlation between infection recurrence in the pairs of kidneys was found to be 0.268 with 95% confidential interval of (-0.1854985, 0.7206918).

Key words: Bivariate, copulas, censoring, correlation, optimization, maximization

INTRODUCTION

The study of associations among bivariate and multivariate outcomes is an interesting research area in statistical science. This is because dependence between two random variables is completely described by their bivariate distribution. When the bivariate distribution has a simple form, standard methods can be used to make inference. However based on particular assumptions, one may alternatively create bivariate distributions thereby restricting its use. Unfortunately, these limitations occur very often when working with bivariate discrete distributions and in most cases they allow only for positive dependence or they can have marginal distributions of a given form.

The distribution of the survival times is uniquely determined by its hazard function h(t) = f(t)/S(t) in univariate survival analysis where f(t) = (S(t))/@(t) and S(t) is the survival time at time t. For bivariate survival analysis such a unique representation in terms of hazard functions does not exist. Although, there are several consistent estimators of the bivariate distribution function, an efficient and consistent estimation has proven to be a difficult problem according to Van der Laan (1996). The 2 representations of the bivariate survival distribution that are most commonly used are due to Dabrowska (1988) and Prentice and Cai (1992). Andersen *et al.* (1993) discussed both representations and the nature of their connections exclusively.

Bivariate survival data are encountered in studies on a frequent basis. Bivariate censored data arise for example in twin studies where the age when one of the twins get a disease may be correlated with the age when his or her twin sibling gets the disease. Obviously, one or both of the twins may not get the disease at all, thus different types of censoring are possible. The dependence between the survival times of the two twins may give us information about genetic or environmental influence on the disease. Although, much research in multivariate survival analysis has focused on methods for inference about the marginal survival times there has also been substantial interest in studying dependence structures. A popular topic of research in multivariate survival analysis is nonparametric estimation of the survival function (Dabrowska, 1988; Prentice and Cai, 1992). Although, the bivariate survival function contains information about the dependence structure, it may be difficult however to visualize because of the discreteness of the estimates of the survival function that have been developed to date, this is the focus of this study.

Corresponding Author: A.F. Fagbamigbe, Department of Epidemiology, Medical Statistics and Environmental Health, Faculty of Public Health, College of Medicine, University of Ibadan, Nigeria

Usually, the structures of correlation are unknown. In recent time, a large volume of research has been directed to modeling failure times without a good outright specification of correlation between the paired censored outcomes. According to a popular approach for modeling, multivariate data is the marginal hazard model approach which models the population-averaged of the covariate effects when the correlation between observations are not of interest. Dabrowska (1988) modeled the joint survival function in terms of two conditional hazard functions and a double hazard function, representing the instantaneous rate of double failures at time t given that both components were alive until that time.

Cumulative versions of each of these 3 hazard functions can be estimated using a Kaplan-Meier type estimate. Further properties of this estimator can be seen in Dabrowska *et al*. (1989) and Gill *et al*. (1995).

Modeling of the bivariate survival function in terms of the two marginal cumulative hazard functions and a covariance rate function was done by Prentice and Cai (1992), it measures the covariance between 2 martingales defined in terms of counting processes. Both representations have attractive properties such as the direct link to the Kaplan Meier estimate (Dabrowska *et al*., 1989) or the relation to the estimation of dependence parameters and the possibility of combining the bivariate model with a marginal proportional hazards model (Prentice and Cai, 1992).

However, neither of them is well suited for estimation with a polynomial spline model since, there are no explicit conditions under which the double hazard function (Dabrowska *et al*., 1989) or the covariance rate function (Prentice and Cai, 1992) yields a valid joint survival function. As a result, the optimization problem has complicated constraints and numerous iterations which are hard to handle also, the corresponding log-likelihood is non-concave even if none of the measured data are censored.

Neither representation yields a direct estimate of the bivariate density. While it is possible to obtain an estimate of the bivariate density as the derivative of the convolution of any estimate of the bivariate distribution function with a kernel, we know of no itch free methodology other than the one presented in this study for directly estimating a correlation of a bivariate density when some data are censored. This study modelled a means of measuring a possible relationship between a set of bivariate survival times which could be censored or not.

Censored observations: In general, censored observations arise whenever the dependent variable of interest represents the time to a terminal event and the duration of the study is limited in time. A special source of difficulty in the analysis of survival data is the possibility that some individuals may not be observed. Suppose that in the absence of censoring, the ith individual in a sample of n has failure time T_i, a random variable.

Also, let c_i be the period of observation such that observation on that individual ceases at c_i if failure has not occurred by then. Then the observation consists of $X_i = \min(T_i, c_i)$ together with the indicator variable:

$$\delta_i = \left\{ \begin{array}{l} 1 \text{ if } T_i \le c_i \text{ (life time is observed)} \\ 0 \text{ if } T_i \ge c_i \text{ (life time is consored)}; i = 1,\ldots,n \end{array} \right\}$$

c_i of individuals who in fact are observed to fail. There are many forms of censoring including type I censoring in which c_i are equal, $c_i = c$, a constant under the control of the investigator and type II censoring where observation ceases after a predetermined number of d failures so c_i becomes a random variable, other forms of censoring exist.

Copula models of joint survival analysis: Copula models have been in use for a long time in estimating association or level of dependency between two censored outcomes though it has its own short comings. Estimation of the joint survival function $S(t_1; t_2)$ may be carried out in two stages. At the first stage, the marginal survival functions $S_1(t_1)$ and $S_2(t_2)$ are estimated non-parametrically (via Kaplan-Meier method or its modifications) or with the help of a parametric model (popular choices could be Gompertz or Weibull distributions). At the second stage, the estimated survival functions are mixed according to some copula models and the association parameter is estimated via maximum likelihood method.

An alternative approach consists in simultaneous estimation of parameters of the marginal distributions and the association parameter. If there is a certain information available regarding the marginals e.g., if there exist additional data available on the individual lives, it is natural to consider Bayesian methods of estimation with informative priors on the marginal parameters and a weak or non-informative prior on the parameter of association. If the second approach is taken and the bivariate survival function is estimated directly as a copula:

$$S(t_1;t_2;\lambda_1;\lambda_2;\rho) = C(S_1(t_1;\lambda_1);S_2(t_2;\lambda_2);\rho) \quad (1)$$

Where:

λ_j = May be vector parameters of the marginals
ρ = The parameter of association

In the presence of right censoring, the ikelihood function for the vector parameter $\lambda = (\lambda_1, \lambda_2, \rho)$ with bivariate survival sets of observation on individual x_i may be represented as:

$$\ell(\lambda|x)=\prod_{i=1}^{n} f(x_{i1}, x_{i2}; \lambda_1, \lambda_2)^{c_{i1}c_{i2}} f_1(x_{i1}, x_{i2}; \lambda_1, \lambda_2)^{c_{i1}(1-c_{i2})} \times f_2(x_{i1}, x_{i2}; \lambda_1, \lambda_2)^{(1-c_{i1})c_{i2}} S(x_{i1}, x_{i2}; \lambda_1, \lambda_2)^{(1-c_{i1})(1-c_{i2})} \quad (2)$$

$$f_j(x_{i1}, x_{i2}; \lambda_1, \lambda_2)=\frac{\partial}{\partial_{xj}} S(x_{i1}, x_{i2}; \lambda_1, \lambda_2)^{-}$$

Where:

x_{ij} = $a_{ij} + t_{ij}$
a_{ij} = The time at which individual
I = Entered the study
t_{ij} = The observation for ith individual
c_{ij} = Censoring indicators

One of attractive model choices considered by Shemyakin and Youn (2006) is the choice of two-parameter Weibull distribution:

$$S(t_j) = \Pr(X_j > t_j) = \exp\left[-\left(\frac{t}{\beta_j}\right)^{\gamma j}\right], t \geq 0 \quad (3)$$

with scales β_j and shapes γ_j, j = 1, 2 for both marginals and stable (Gumbel-Hougaard) copula for the model of association. The resulting copula representation is:

$$S(t_1;t_2)=C_{GH}(S(t_1;\beta_1;\gamma_1);S(t_2;\beta_2;\gamma_2);\alpha) = \exp\left(-\left[\left(\frac{t}{\beta_1}\right)^{\alpha\gamma 1}\left(\frac{t}{\beta_2}\right)^{\alpha\gamma 2}\right]^{1/\alpha}\right) \quad (4)$$

maximum likelihood estimation for this model has been carried out by Frees and Valdez (1998) where Frank's copula and Gompertz marginals were also considered. Bayesian estimation with exponential and normal priors for Weibull distribution parameters, Beta prior for the association and the copula functions of three forms: stable, Frank's and Clayton's was performed. This approach is good and may be very effective. However, it sometimes proves to be insufficient depending on the type of association between the paired lives. The source of this insufficiency, however is the problem of dimensionality. For instance, the use of two-variate survival function $S(x_1; x_2)$ in order to model the behavior of three-variate joint first-life and last-survivor functions. The applications of bivariate copula models for discrete data are limited. Usually, there is need to trade off between models with limited dependence only positive association and models with flexible dependence but computational intractability. For an example, the elliptical copulas provide a wide range of flexible dependence but do not have closed form cumulative distribution functions. So, one needs to evaluate the bivariate copula and hence, a bivariate integral repeated for a large number of times. This can be time consuming but also because of the numerical approach used to evaluate a multivariate integral, it may produce round off errors.

On the other hand, bivariate Archimedean copulas, partially-symmetric m-variate copulas with m-1 dependence parameters and copulas that are mixtures of max-infinitely divisible bivariate copulas have closed form cumulative distribution functions and thus computations are easy but allow only positive dependence among the random variables. The bridge of the two above mentioned problems might be the definition of a copula family which has simple form for its distribution function while allowing for negative dependence among the variables. A possible way out might be to define such a copula family exploiting the use of finite mixture of simple uncorrelated normal distributions.

Since the correlation vanishes, the cumulative distribution is simply the product of univariate normal cumulative distribution functions. The mixing operation introduces dependence. Hence, we obtain a kind of flexible dependence which could be negative. This is the problem that motivated this research. To my knowledge, there is no formal work elaborating or providing a viable alternative to this problem in literature. It is highly necessary and of course important to develop a flexible and effective universal estimation method of this type of association. The objective of this research is to model an alternative method as a result of deficiencies of the copula models. This study is aimed at modeling a possible association between censored paired items using the inverse of Kapler-Miere estimates which are basically non-parametric. A kidney infection recurrence data gotten from experiment conducted on 38 patients in North Eastern part of England is used as the case study.

MATERIALS AND METHODS

Data: A kidney infection recurrence data gotten from experiment conducted on N = 38 patients in North Eastern part of England was used as the case study. Each N patient were observed and times between recurrences of a particular type of event are recorded. The problem that motivated the experiment is the recurrence of infection in kidney patients who are using a portable dialysis machine. The occurrence of the infection begins at the point of

insertion of the catheter and when it occurs, the catheter must be removed; the infection cleared up and then reinserted. Recurrence times are times from insertion until the next infection. Sometimes, the catheters are removed for other reasons so that there may be a right censoring of the data. Also, the final reoccurrence time may be censored. It is assumed here that each patient is followed for a predetermined number of reoccurrence times, some of which may be censored. The recurrence ties for each individual have a common frailty that may be regarded as a random selection from suitably defined population distribution of frailties. This is as a result of 10 weeks interval allowed between an infection and reinsertion of the catheter, the recurrence intervals are taken to be independent apart from their common frailty component. The variables recorded for each individual are:

i = Patient number
T_{ij} = jth smallest recurrence times for patient i
X_{ij} = Vector of risk variables applying to patient i at his jth ordered reoccurrence time
c_{ij} = Censoring indicator

$$\delta_i = \begin{cases} 1 \text{if lifetime is observed} \\ 0 \text{if lifetime is censord}; i = 1,2 \end{cases}$$

The risk variables are age, sex (1 = male, 2 = female) and disease type coded as 0 = GN, 1 = AN, 2 = PKD and 3 = others. A portable dialysis machine is a portable kidney dialysis machine that allows an in-home treatment, enabling hundreds of thousands of people afflicted with kidney failure to treat themselves at home instead of traveling to dialysis clinics 3 days a week.

Data analysis: The study relies on the principle of optimization of likelihoods of the event of interest occuring. The four possible scenarios in bivariate censored data were considered. A particular case followed over time may have neither type of relapse (a good outcome), one or the other or both. However, all the outcomes were subjected to censoring.

The Kaplan-Meier estimates of the survival times of each univariates were obtained and transformed to normal data using the inverse of the estimates as shown in Eq. 5. This was with a view of transforming a non-normal data to data under normal distribution, taking into account the censoring of the observations. This was used as a remedy for outliers, failures of normality, linearity and homoscedasticity and also to:

$$\Phi^{-1}\{F_{km}(t_i)\} \quad (5)$$

Produce a data that is approximate to bivariate standard normal distribution. This is appropriate since, Kaplan-Meier estimator is simply a staircase function with the location of the drops randomly placed and due to censoring, the size of the drops also changes (increases as time increase or as censoring counts increases).

Thus for each observation, there is 50% chance of been censored. The censoring times are identical and independent exponential random variables. Hence, there is need to invert the estimates so as to get a normalized data. In an attempt to check the procedure of data transformation under the inversion of Kaplan-Meier estimates, one of the conventional data transformation methods logarithm transformation [log(x)] is used as a control. The results are very similar as shown in Table 1. The transformed data was later standardized using:

$$x_i^* = \frac{\text{Mean}(x) - x_i}{\text{Standard Deviation}(x)} \quad (6)$$

which gave a desired normal random variable X and Y with mean zero and variance one where X and Y are both censored and ρ is the correlation coefficient of X and Y Kolmogorov-Smirnov test was used for comparison of the two data sets and to check for the normality of the transformed data.

We engaged Quasi-Newton method, an optimization technique of optimization otherwise called maximization with the aim of finding the position θ which maximizes the function $L(\theta)$.

The function could be either likelihood or loglikelihood. It involves taking routine steps of varying length and direction, around the parameter space until convergence is deemed to have been achieved. It is a general purpose optimization model based on Nelder-Mead, Quasi-Newton and conjugate-gradient algorithms. It includes an option for box-constrained optimization and simulated annealing.

The function optimize is used instead of optim used in this study because the present study considered only one-dimensional parameter, it searches the interval from lower to upper for a minimum or maximum of the function f with respect to its first argument.

Table 1: Kolmogorov-Smirnov test for Kaplan-Meier and Logarithm transformations

	Kaplan-Meier transformation		Logarithm transformation	
Statistics\variables	X	Y	X	Y
p	0.87	0.68	0.72	0.92
Mean	4.3767E-02	-2.1313E-02	1.823E-01	4.6302E-03
Variance	0.8921	0.9100	1.008016	1.065024

Limitations: One of the known shortcomings in this modelling is the use of marginal survival functions in determining the Kaplan-Meier estimates rather than joint survival function because Kaplan-Meier researches by ordering the survival time marginally thereby creating the task of re-connecting each pair after the normal transformation.

Other bottlenecks encountered in this study is during the normal transformation where it became unavoidable to introduce a fiddle factor so as to get rid of infinity values at the two tails of the resulting transformed normal data. The qnorm (1) = 1.

The simplest way is to put in a fiddle factor (1 or 2) that removes that infinity. It is not completely arbitrary in that the smallest transformed value you get is qnorm (1/n) where n is the dataset size. Therefore, the 1-quantile of the first value is roughly the quantile of the last value.

Bivariate normal density: Let X and Y be two random variables normally distributed with mean (μ_x, μ_y) and variance (δx , δy). Also, let ρ be correlation coefficient between X and Y. The probability density function f(x, y; μ_x, μ_y, δx , δy, ρ) is:

$$\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{(x-\mu_x)}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right) \quad (7)$$

In this study, X and Y were standardized variates with zero means and unit variances. Equation 8 reduces to:

$$f(x,y;\rho)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right) \quad (8)$$

Bivariate density estimation for censored data: Let consider n randomly selected pairs $T_i = (X_i; Y_i)$, $1 \le i \le n$. It is assumed that T_i has density f(.) and that the marginal densities of X_i and Y_i are normal on [0; 1]. For $1 \le i \le n$ and $l \in (1, 2)$, let C_{li} be the censoring time for the lth component of the ith pair ($C_{li} = 1$ if this component is uncensored). Set $T_{li} = \min(T_{li}; C_{li})$ and _li = ind (X_{li} _ C_{li}).

Also set $T_i = (X_i ; Y_i)$, $C_i = (C_{1i} ; C_{2i})$ and i = (1i; 2i). It is assumed that (X, Y) and C are independent. The random variable T is said to be uncensored if $\delta = (1; 1)$, censored in the first component only if $\delta = (0; 1)$, censored in the second component only if $\delta = (1; 0)$ and doubly censored if $\delta = (0; 0)$.

Integration: Most if not all, bivariate integrals can be reduced analytically to univariate integrals because linear splines are used to model the logdensity. The resulting univariate integrals are closely related to the exponential integral (Abramowitz and Stegun, 1965) for which fast algorithms exist. This is very important when some data may be censored since, integrals have to be computed for every censored observation. If cubic splines were used to model the density, it would no longer be possible to reduce the bivariate integrals to univariate ones. It is worth noting that if cubic splines were used to model the density, it would be highly necessary to first transform the data to avoid complications because of tail constraints (Koo, 1996).

After the data have been transformed, optimization is used to find the maximum likelihood estimate. Although, the log-likelihood function is not necessarily concave when some data are censored, no major numerical difficulties has been experienced during the computations.

A means of getting fast code is the efficient organization of the computation of the score function and the Hessian involving the integrals discussed above. This situation is in relation to the experience of Kooperberg and Stone (1992) in the context of univariate logspline density estimation. In fact, the real and simulated data sets both took <7 sec of CPU time on the R language used. The likelihood corresponding to different categories is as expressed as:

The likelihood: Contributions of the four possible scenarios of censored bivariate normal variables to the likelihood. For the data set with possibility of censoring both variables, the likelihood consist of four different scenarios.

1st scenario: Let $H_1(x, y)$ denote the likelihood of the situation where the two life times are observed, Then:

$$H_1(x,y)=f(x,y;\rho) \quad (9)$$

where, f(x, y; ρ) is the usual bivariate normal distribution function.

2nd scenario: Let H_2 denote the likelihood of the situation where the first lifetime is observed (X) and second lifetime (Y) is censored at c_2:

$$\begin{aligned} H_2(x,y) &= \int_{c2}^{\infty} f(x,y;\rho)dy \\ &= \int_{c2}^{\infty} f(y|x;\rho)f(x)dy \\ &= f(x)\int_{c2}^{\infty} f(y|x;\rho)dy \\ &= f(x)\left[1 - F_{Y|X=x}(c_2)\right] \end{aligned} \quad (10)$$

where, $F_{Y|X=x}$ is normally distributed with mean:

$$\mu_{Y|X=x} = \mu Y + \frac{\rho\sigma Y}{\sigma X}(x - \mu X) = \rho x$$

and variance:

$$Var(Y - X = x) = \sigma^2 Y(1-\rho^2) = 1-\rho^2$$

Also:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{x-\mu_x}{\sigma_x}\right)^2 = \frac{1}{\sqrt{2\pi}}\exp(x^2)$$

3rd scenerio: Let $H_3(x, y)$ denote the likelihood of the situation where the second lifetime is observed (Y) and first lifetime (X) is censored at c_1.

$$\begin{aligned} H_3(x,y) &= \int_{c1}^{\infty} f(x,y;\rho)dx \\ &= \int_{c1}^{\infty} f(y|x;\rho)f(y)dx \\ &= f(y)\int_{c1}^{\infty} f(x|y;\rho)dx \\ &= f(y)\left[1 - F_{X|Y=y}(c_1)\right] \end{aligned} \quad (11)$$

4th scenerio: Let $H_4(x, y)$ denote the likelihood of the situation where the two lifetime are both censored:

$$\begin{aligned} H_4(x,y) &= \int_{c1}^{\infty}\int_{c2}^{\infty} f(x,y;\rho)dxdy \\ &= 1 - F_X(C_1) - F_Y(C_2) + F_{X,Y}(C_1, C_2;\rho) \end{aligned} \quad (12)$$

Where, $F_X(c_1)$ is the normal cumulative distribution function of X at c_1 and $F_{X,Y}$ (c_1, c_2) is the bivariate cumulative distribution funtion of X and Y at c_1 and c_2, respectively with correlation ρ.

Likelihood estimation: The likelihood for jth observation is then computed from Eq. 9-12 to give:

$$L_j = H_1^{\delta_1\delta_2} H_2^{\delta_1(1-\delta_2)} H_3^{(1-\delta_1)\delta_2} H_4^{(1-\delta_1)(1-\delta_2)} \quad (13)$$

for j = 1,....., n where:

$$\delta_i = \begin{cases} 1 & \text{if lifetime is observed} \\ 0 & \text{if lifetime is censored;} \quad i = 1,2 \end{cases}$$

and log-likelihood for the jth observation is given by:

$$\begin{aligned} \log L_j = {} & \delta_1\delta_2 \log H_1 + \delta_1(1-\delta_2)\log H_2 + \\ & (1-\delta_1)\delta_2 \log H_3 + (1-\delta_1)(1-\delta_2)\log H_4 \end{aligned} \quad (14)$$

Therefore, the overall likelihood is:

$$\ell = \log L = \sum_{j=1}^{n}(\log L_j) \quad (15)$$

Where, n is the number of individuals. Let $\hat{f}_1$ and $\hat{f}_2$ be estimates of the marginal densities of X and Y, respectively and let $\hat{F}_1$ and $\hat{F}_2$ be the corresponding estimated distribution functions. Set $T_i^* = \hat{F}_j(T_{1i}), 1 \in (1,2)$ and apply the procedure described in the previous subsection to $T_i^* = T_{1i}^* T_{2i}^*$ and δ_i with $1 \le i \le n$ to obtain an estimate of the density of (X, Y)* = ($\hat{F}_1$ (X); $\hat{F}_2$ (y)) (X, Y) = ($\hat{F}_1$ (X); $\hat{F}_2$ (y)). The estimate of the bivariate density of (X,Y) is then given by:

$$\hat{f}(.) = \hat{f}^*(\hat{F}_1(x), \hat{F}_2(y)))\hat{f}_1(x)\hat{f}_2(y), 0 \le x \le y \le \infty$$

The estimates of the distribution function F(t) and survival function S(t) = P(X≥x, Y≥y) are:

$$\hat{F}(x,y) = \int_{-\infty}^{\hat{F}(x)}\int_{-\infty}^{\hat{F}(y)} \hat{F}(x,y)dxd\hat{y}F(x,y)$$

and:

$$S(x,y) = \int_{\hat{F}(x)}^{\infty}\int_{\hat{F}(y)}^{\infty} \hat{f}(x,y)dxdy$$

respectively.

RESULTS AND DISCUSSION

The survival function of the recurrence time in each of the two kidneys is as shown in Fig. 1a.

The overall average recurrence time and its standard deviation are 102 and 131, respectively though the recurrence time in the first kidney has average and standard deviation of 112 and 144.01, respectively while the average and standard deviation of recurrence time in the second kidney recurrence time is 92 and 117.20, respectively. The plot also showed that the mode of the recurrence time in the two kidneys is 42. Figure 1b shows the Kaplan-Meier curve of the survival function $S(t_1, t_2)$ using the recurrence time of each observation in the study.

Stratification of explanatory variables: An attempt to stratify the response by sex and disease types so as check if there is any difference between each explanatory variable (Fig. 2) showed that kidney infection differs

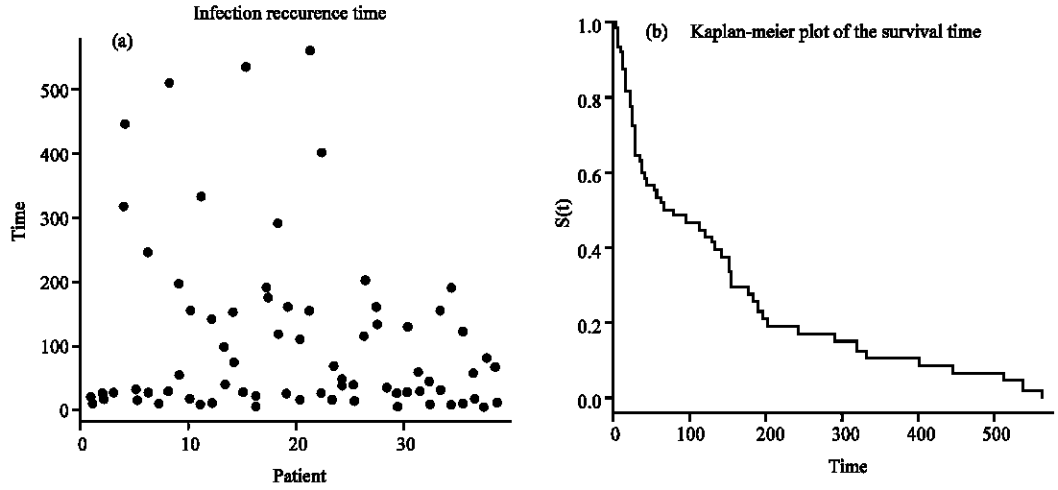

Fig. 1: Patient to recurrence time and Kaplan-Meier curve of the recur-rence time

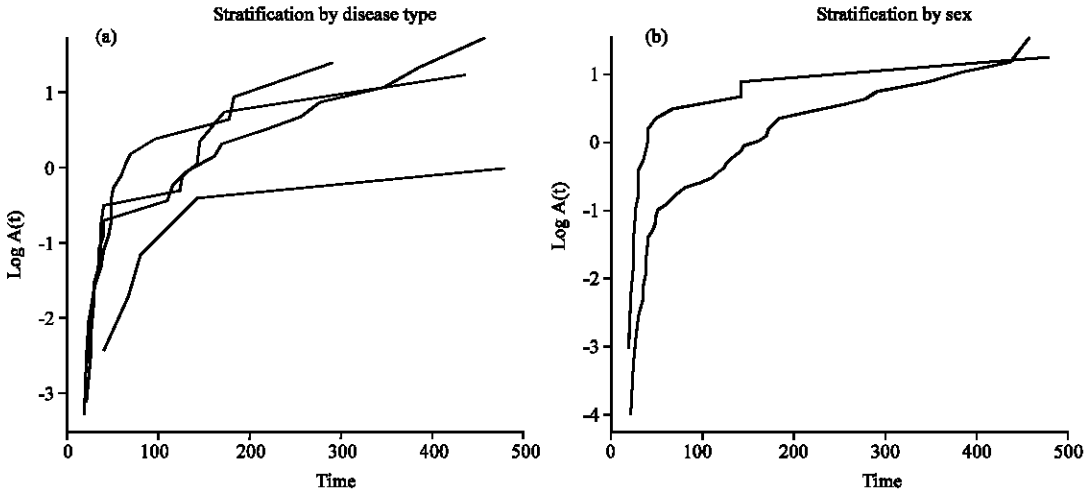

Fig. 2: Stratification of the variables

between male and female and also that different types of diseases results in different level of kidney infection.

Prognostic indexes: The plot of the survival lines using the available prognostic indexes is as shown in Fig. 3. In each plot, the observed survival line is shown by block lines while the fitted survival lines are shown in dash lines.

The prognostic indexes are factors that can possibly predict, explain or affect the disease (response) under investigation, here i have considered age, sex and disease as prognostic indexes, the plot in the centre above shows the observed and fitted survival line by age and disease type as prognostic indexes while the right-most graph depicts the observed and fitted survival line where the prognostic indexes are sex and disease types.

Survival functions for the recurrence times: Figure 4 is the graph of the survival function S(t) for both set of observations. Figure 4 showed that there are differences between the two set of responses with the survival function of the first set of observations $S(t_1)$ falling more steadily than the survival function of the second observation $S(t_2)$.

Similar to Fig. 4, the cumulative density functions for the two set of recurrence time showed that there is significant difference between the two with the cumulative density function (cdf) of the first set of observations $F(t_1)$ rising more sharply than the cdf of the second observation $F(t_2)$. This is usual as F(t) and S(t) are inverse of each other as defined by relationship:

$$F(t_i) = 1 - S(t_i)$$

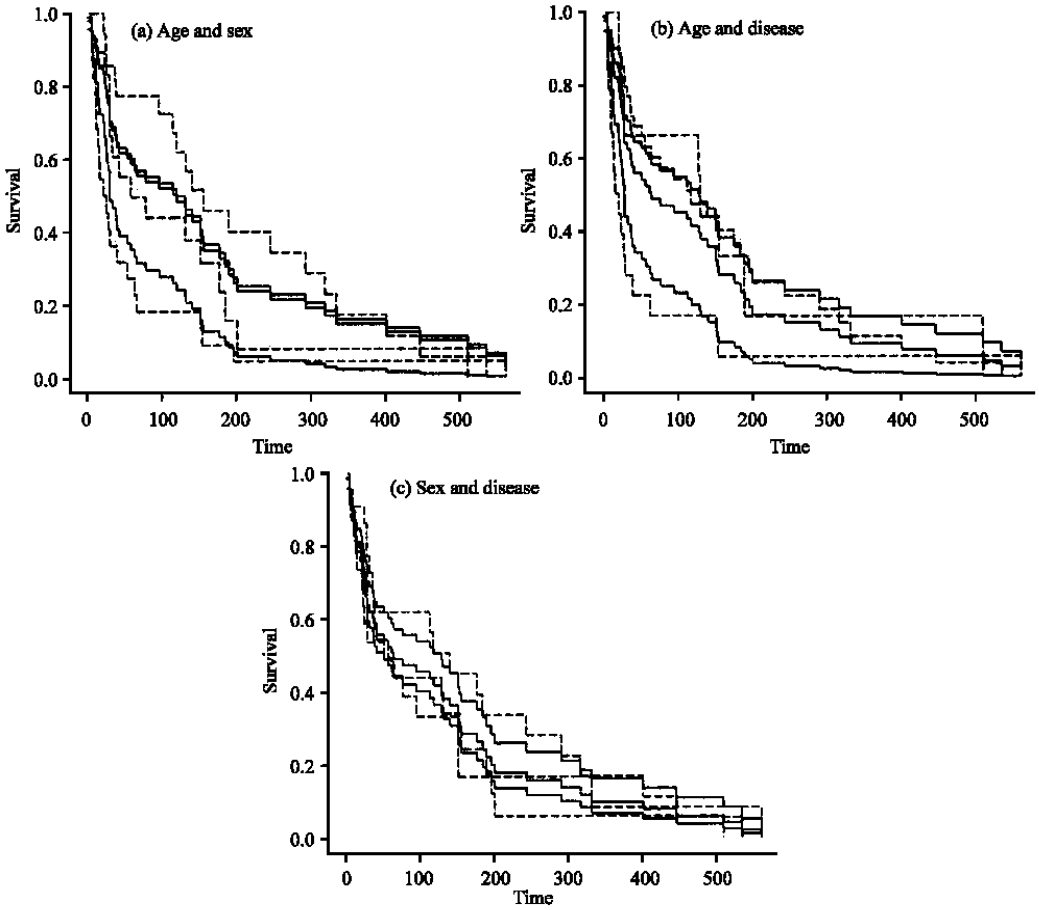

Fig. 3: Prognostic indexes by sex, age and diseaese types. Solid line is fitted and dotted is observed

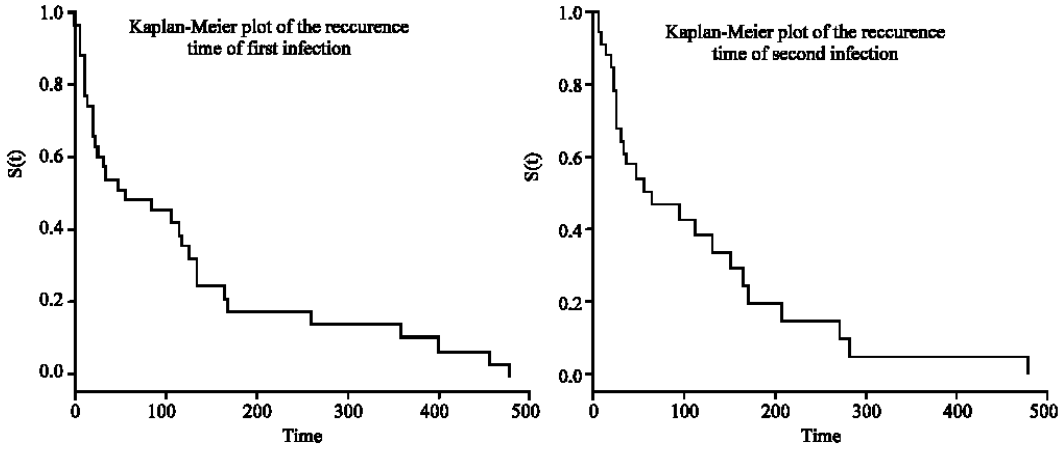

Fig. 4: Survival functions for the two recurrence times

Normalisation and optimization: Table 1 shows the result of Kolmogorov-Smirnov test used to check the normality of the resulting data. It was found out that the transformed data is consistent with data under normal distribution. Figure 5 showed clearly that the transformed data are normally distributed. The same test was also used to check the data i obtained using logarithm transformation. The outcome is shown in Table 1. The optimization procedure using the function optimize instead of optim because the default method may not

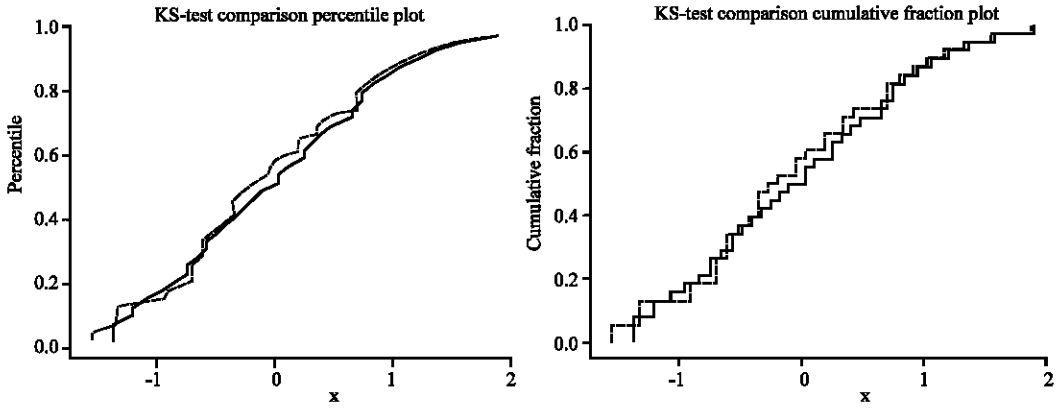

Fig. 5: Kolmogorov-smirnov test on the transformed recurrence times Using inverse of Kaplan-Meier and logarithm transformation methods

work well and give warnings. However, the study showed that the two functions optim and optimize gave same results with successful convergence in the two R-codes.

As shown in Table 2, the correlation between recurrence time of kidney infection is 0.268 with 95% confidence interval of (-0.1854985, 0.7206918). The Kolmogorov-Smirnov test used to check if the resulting data is normally distributed or not found that the data (X and Y) is consistent with a normal distribution.

Also, a similar test used in checking the normality of resulting data under the logarithm transformation also showed that the data (X and Y) is consistent with a normal distribution having the parameters shown before.

Using the optimization procedure, the correlation between recurrence time of kidney infection is 0.268 with 95% confidence interval of (-0.1854985, 0.7206918) using the transformation of the inverse of the Kaplan-Meier estimates which is very close to the estimate of 0.2789395 with 95% confidence interval of (-0.05022155,0.6081005) when the logarithm data transformation technique was used.

The wide confidence interval is as a result of the relatively small sample size of 38 used for the study. The result simply showed that are dependence among the recurrence time of kidney infections and in fact, there are about 27 in 100 chances of having a recurrence of infection in the two kidneys. Ordinarily a correlation coefficient of zero signifies independency (no correlation) and it will be noticed that the association parameter has a confidence interval (-0.051, 0.608) which actually has zero within the interval. A larger sample size can change the limit of the confidence interval and thereby excluding zero and hence makes the association between the bivariate outcomes significant.

Table 2: The maximum likelihood estimate of the correlation between the recurrence times

Transformation method	Inverse of Kaplan-Meier	Logarithm
Likelihood	83.3164200	90.7252700
Correlation coefficient (p)	0.2675799	0.2789395
Standard error of p	0.2311710	0.1679393
CI for p at 95%	(-0.1855, 0.7207)	(-0.0502,0.6081)

CONCLUSION

The study examined the four different situations that can arise in a bivariate censored data. It is based on usage of inverse of Kaplan-Meier estimates (a non-parametric estimate) to get a transformed normal data approximate to standard normal data because the only parameter of interest is the association parameter. Other simpler and common methods of Normal data transformation were discussed. Kolmogrov-Smirnov test was used to ascertain the normality of the transformed data.

A particular case followed over time may have neither type of relapse (a good outcome), one or the other or both. However, all the outcomes will be subjected to censoring.

Contributions to the likelihood of the four different scenarios that can arise in a bivariate censored data were examined critically. The result of the maximization of the likelihoods is the desired association parameter.

The logarithm transformation of Kaplan-Miere estimates to normally distributed data seem to have worked on this occasion though this is not guaranteed in all cases.

The study also found that disease types and sex are the only significant variables affecting kidney infection. The model used in this study will work efficiently in all bivariates outcomes whether they are subjected to censoring or not if the prescribed procedures are adequately adhered to and large and unbias

representative samples are used . It will work not only within health and medical research but also in social sciences, insurance and financial studies.

REFERENCES

Abramowitz, M. and I.A. Stegun, 1965. Handbook of Mathematical Functions. Dover Publications, New York.

Andersen, P.K., Q. Borgan and R.D. Gill, 1993. Statistical Models Based on Counting Processes. Springer, New York.

Dabrowska, D.M., 1988. Kaplan-meier estimate on the plane. Ann. Statist., 16: 1475-1489.

Dabrowska, D.M., K.A. Doksum and J.K. Song, 1989. Graphical comparison of cumulative hazards for two populations. Biometrika, 76: 763-773.

Frees, E.W. and E.A. Valdez, 1998. Understanding relationships using copulas. North Am. Actuarial J., 2: 1-25.

Gill, R.D., M.J. van der Laan and J.A. Wellner, 1995. Inefficient estimators of the bivariate survival function for three models. Annales de l'institut Henri Poincar (B) Probabilits et Statistiques, 31: 545-597.

Koo, J.Y., 1996. Bivariate B-splines for tensor logspline density estimation. Comput. Statist. Data Anal., 21: 31-42.

Kooperberg, C. and C.J. Stone, 1992. Logspline density estimation for censored data. J. Comput. Graph. Statist., 1: 301-328.

Prentice, R.L. and J. Cai, 1992. Covariance and survival function estimation using censored multivariate failure time data. Biometrika, 79: 495-512.

Shemyakin, A.E. and H. Youn, 2006. Copula models of joint survival analysis. Applied Stochastic Models Bus. Industry, 22: 211-224.

Van der Laan, M.J., 1996. Efficient and inefficient estimation in semiparametric models. Technical Report, CWI, Amsterdam.