**A Survival analysis model for measuring Association between Bivariate Censored Outcomes: Validation Using Mathematical Simulation**

1Fagbamigbe AF\*, 1Adebowale AS, 1Bamgboye EA

\*Corresponding Author (franstel74@yahoo.com, +2348061348165)

1Department of Epidemiology and Medical Statistics, Faculty of Public Health, College of Medicine, University of Ibadan, Ibadan, Nigeria

Abstract

Bivariate censored data occur in follow-up studies of diseases that can result in two different outcomes. Many studies have explored methods for inference about the marginal recurrence times of these outcomes.However, very few have focused on the dependence structures between their recurrence times as evidence in the current study. We reviewed and validated the earlier developed model for measuring the association.

This theoretical and empirical study used simulated data to validatethe survival analysis model for measuring association between recurrence times of bivariate censored outcomes. Using predetermined correlation coefficients, two sets of n=1000 standardized binormal data were simulated at 0% and 50% censoring. A particular datum from the standardized binormal data would naturally fall into one of four possibilities: only the first, only the second, none or both conditions occurring with likelihoods H1, H2, H3 and H4 respectively.Contributions of the likelihoods were examined and used with standardized binormal data to obtain the maximum likelihood estimate of the association parameter between the outcomes.

For the data simulated at 50% censoring, maximum likelihood estimates of the association parameter tended to zero as the predetermined correlation coefficients fell from +1.0 to -1.0. However, at 0% censoring, the maximum likelihood estimates were approximates of the predetermined correlation coefficients.

The developed model was robust as maximized values varied considerably with different transformation methods. It would be useful in studying associations between two censored recurrence times of diseases in similar studies.

Keywords: Bivariate Censored outcomes; Maximum likelihood estimates; Censoring; Simulation

**INTRODUCTION**

Bivariate censored data arise in the study of a phenomenon with two possible outcomes in which the time before occurrence or recurrence of the event of interest may be the same or differ in the two outcomes[1]. It is of interest to determine, if it exists, the possible associations between the times of occurrence of these events especially if both were subjected to censoring. Although, previous studies in bivariate survival analysis have focused mainly on methods for inference about the marginal survival times, very few have investigated dependence structures between bivariate outcomes [2–5] as found in this study. The study objective is to determine the dependence structure of the timing of the variates with the view of providing insight into whether occurrence of one of the outcomes depends on the other or not.

Research in multivariate survival analysis which had been widely explored is non-parametric estimation of the survival function[4–6].Deviation from this research is the bivariate survival function which has information about the dependence structure. Limited study on this is often because the dependence structure is hard to visualize due to the discreteness nature of the developing method of estimating its survival function.

The study of associations among bivariate and multivariate outcomes becomes a necessity in scientific research because dependence between two random variables is completely described by their bivariate distribution. The well-known standard methods could be used to make inferences when a bivariate distribution has a simple form but constraints usually surface when the form is a bit complex. However, based on some standardized assumptions, one may alternatively create bivariate distributions thereby restricting its use. These limitations occur very often when working with bivariate discrete distributions and in most cases they allow only for positive dependence or have marginal distributions of a given form[7].

In some biomedical applications of survival analysis techniques, interest focuses on the dependence relationship between the lifetimes of two variables in an individual. For example, the analysis of data on lifetimes of twins has been used by geneticists as a tool for assessing genetic effect on mortality[8].Also, in the studies of HIV/AIDS, the dependence between the time from HIV infection to full blown AIDS and the time from full blown AIDS to death have revealed useful information about the evolution of the disease process.

The study of the estimation of Kendal’s τ under censoring showed that censoring is a common phenomenon in analysis of lifetime data and it is essential that estimates of τ be available for bivariate censored data[9]. However, few results for this fundamental problem have appeared in the literature. Several estimators of τ under censoringhave been proposed [10–12], but none of the estimators is consistent when the true value of τ is not equal to zero, that is, when the marginals are dependent. The bias of these estimators increases as the degree of dependence increases. They had expressed τ as an integral of the bivariate survival function. Adopting the ideas of [13]a natural way to estimate τ is to plug a suitable bivariate survival estimator into the integral form that defines τ [9].

In a study that assessed the association for bivariate current status data[14], utilized interval censoring rather than right censoring that is the commonest type of censoring and the concept were utilized in this study. Wang and colleague proposed a method for measuring correlations in the presence of interval censored data. This was adopted in the study of characterization of the correlation between ages at entry into Breast and Pubic Hair development[15]. Interval censoring occurs in a survival data situation when its univariate setting is considered and one is interested in a fatigue time variable T which is never observed, but can only be determined to lie below or above a random monitoring or censoring time C where C and T are assumed to be independent[16].

Copula models is one major approach used in estimating association between bivariate censored data but there has been criticism against the method.It was stated that the approach is elegant and may be effective, it sometimes proves to be insufficient depending on the type of association between the paired observations [17].The source of this insufficiency, apparently, is the problem of dimensionality, where a two-variate survival function S(x1; x2)is used in order to model the behaviour of three-variate joint first-life and last-survivor functions. It’s over reliance on Bayesian techniques by using prior distributions at the second stage in copulamodelsmakes the method a cumbersome. Hence the parametric modelling of the association was developed in a previous study as a simple model that less relied on Bayesian techniques [18]. Where, a model for determining existence and measuring association between bivariate censored outcomes was formulated. The study illustrated the model using censored outcomes of recurrences of kidney infections. They found that positive and relatively strong associations exist between recurrence times in kidney infection and was optimized at +0.2868.

The present study primarily attempted to unpack and examine the tight mesh of ideas underpinning modeling an association between bivariate outcome using data generated via statistical simulations and to highlight dimensions that have been neglected in our previous study.Our major focus was validation of the procedures of the model developed in [18]using a set of predetermined correlation coefficients at 0% censoring (when all events are presumed to be observed) and 50% censoring (when all events haveequal chances of beingobserved or not).

**Methodology**

Using R statistical software, we simulated sets of *n=1000* standardized binormal random variables -*(x, y) ~ mvn(0, 1; 0, 1)* with a set of predetermined correlation coefficients (ρ) from -1.0, to +1.0 where *mvn* is a syntax in R-statistical package for simulating multivariate random variable with a specified mean (0) and standard deviation (1). Each simulated data were in turn assigned to two different censoring conditions. Under the first condition, each bivariate simulated data was randomly assigned censoring index with equal chances (50%) so that each has either index *1 or 0;* for observed and *0* for censored observations. However, each data in the two marginals of the simulated data were all assigned 0% censoring index indicating that all events were observed under the second condition. We then applied the simulated data with their respective censoring index to the model $l = log L=\sum\_{j=1}^{n}log L\_{j} $[18] and as detailed in equations (1) to (8) below so as to obtain the Maximum Likelihood Estimation of association parameter for each of the data (under 0% and 50% censoring). We used normal multivariate distribution estimators (pnorm, dnorm, pnorm, pmvnorm and dmvnorm) in R-statistical package to integrate and determine the necessary functions. Details already published [18] To ensure the reliability and validity of the simulated data, we computed its theoretical correlation coefficient using Pearson’s product moment correlation methods. Hessian matrix was used in computing the standard error of the maximum likelihood estimations.

It has been suggested that in bivariate survival analysis that two outcomes (observations) occur in any follow up study[2,14,18].A particular case followed over time may have neither of the two events of interest,(a good outcome), one or the other, or both (a bad outcome). The likelihoods corresponding to the four different categories are expressed individually as *H1(x, y)* which involves a situation when the two outcomes, say *X* and *Y* are both observed during the follow up. However, the survival (observational) time of the two outcomes may not be the same. Let *H2(x,y)* denote the likelihood of the situation where the first lifetime is observed *(X)* and second lifetime *(Y)* is censored at C2, while *H­3(x,y)* denote the likelihood of the situation where the first lifetime *(X)* is censored at C1and second lifetime *(Y)* is observed. *H4(x,y)* denotes the fourth possible situation whereby neither of the events of interest took place but the survival time was rather censored, that is let *H4(x, y)* denote the likelihood of the situation where the two lifetimes *X* and *Y* are censored at C1 and C2 respectively. The censoring time C1 and C2may or may not be the same.

Then, the overall likelihood model to be maximized could be summarized as

$l = log L=\sum\_{j=1}^{n}log L\_{j}$**……………………………………………………..………….(1)**

where

$L\_{j}=H\_{1}^{δ\_{1}δ\_{2}}H\_{2}^{δ\_{1}(1-δ\_{2})}H\_{3}^{(1-δ\_{1})δ\_{2}}H\_{4}^{(1-δ\_{1})(1-δ\_{2})}$**……………………………..………………(2)**

and

$$log L\_{j}=δ\_{1}δ\_{2}logH\_{1}+ δ\_{1}(1-δ\_{2})logH\_{2}+ (1-δ\_{1})δ\_{2}logH\_{3}+(1-δ\_{1})(1-δ\_{2})logH\_{4}$$

*for j=1,............,n* **…………………………………………………………………………..(3)**

$δ\_{i}=\{\_{=0 if lifetime is censored}^{=1 if lifetime is observed}$*fori=1,2*,**……………………………….…………..(4)**

$H\_{1}=f(x,y;ρ)$ **……………………………………………………………..…..………….(5)**

$H\_{2}=f(x)[1-F\_{(Y|X=x)}](c\_{2})$ **………………………………………..……………….……(6)**

$H\_{3}=f(y)[1-F\_{(X|Y=y)}](c\_{1})$ **……………………………………………………….……..(7)**

$H\_{4}=1-F\_{X}(c\_{1})-F\_{Y}(c\_{2})-F\_{XY}(c\_{1},c\_{2};ρ)$ **……………………..………………………..(8)**

Also,

*C1*= the censoring time for the first outcome

*C2*= the censoring time for the second outcome

*FX (c1)* is the normal cumulative distribution function of *X* at *c1*

*Fy (c2)* is the normal cumulative distribution function of *Y* at *c2*

*FX|Y=y(c1)* is the marginal distribution function of *X* at *c1* given *Y*

*FY|X=x(c2*) is the marginal distribution function of *Y* at *c2* given *X*

*FX,Y(c1,c2;*ρ) is the bivariate cumulative distribution function of *X* and *Y* at*c1*and *c2* respectively with correlation ρ.

**Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Using 50% censoring chance we generated censoring index for the n=1000 bivariate simulated data and obtained the maximum likelihoods and the corresponding estimate of the association parameter (ρ) at convergencebetween the bivariate simulated data. In Table 1, Column (a) shows the predetermined correlation coefficients used in carrying out the simulation while Column (e) is a check using theoretical Pearson’s product moment correlation coefficient that the simulations were done correctly. The optimization model in equation (1) above as documented in[18] was used to determine the likelihood function of the association parameter (b), the estimate of the association parameter (c) as well as its standard error (d). | a | b | c | d | e |  |
|  | Simulation | Likelihood of | Estimated Rho (ρ) | Standard Error | Theoretical |  |
|  | Rho (ρ) | function(Lj) | at 50% Censoring | se(ρ) | Rho (ρ) |  |
|  | +1.00 | 1051.185 | 0.9999500 | 0.00098733 | 1.0000000 |  |
|  | +0.90 | 1842.248 | 0.9146428 | 0.01256556 | 0.8996096 |  |
|  | +0.80 | 2042.695 | 0.8589264 | 0.01838316 | 0.7955383 |  |
|  | +0.75 | 2051.601 | 0.8330821 | 0.02124831 | 0.7526036 |  |
|  | +0.70 | 2102.807 | 0.7874704 | 0.02486063 | 0.6945606 |  |
|  | +0.60 | 2190.787 | 0.7276363 | 0.02956057 | 0.6033463 |  |
|  | +0.50 | 2230.212 | 0.7108717 | 0.03072535 | 0.4931400 |  |
|  | +0.40 | 2286.953 | 0.6807115 | 0.03122125 | 0.3954632 |  |
|  | +0.30 | 2303.721 | 0.6648109 | 0.03310495 | 0.3052653 |  |
|  | +0.25 | 2357.027 | 0.6635675 | 0.03171611 | 0.2512262 |  |
|  | +0.20 | 2339.695 | 0.6302059 | 0.03470269 | 0.2151449 |  |
|  | +0.10 | 2342.918 | 0.6531314 | 0.03238656 | 0.1084885 |  |
|  | 0.00 | 2442.966 | 0.5442600 | 0.01445169 | 0.0039800 |  |
|  | -0.10 | 2348.513 | 0.5765490 | 0.03780633 | -0.1118956 |  |
|  | -0.20 | 2543.847 | 0.4861451 | 0.03995222 | -0.2157024 |  |
|  | -0.25 | 2584.637 | 0.4859900 | 0.04081206 | -0.2514401 |  |
|  | -0.30 | 2551.681 | 0.4835510 | 0.04009839 | -0.3169049 |  |
|  | -0.40 | 2512.693 | 0.4555082 | 0.04079666 | -0.4039177 |  |
|  | -0.50 | 2528.321 | 0.4931571 | 0.04054079 | 0.4941700 |  |
|  | -0.60 | 2538.477 | 0.4678533 | 0.04012985 | -0.6145968 |  |
|  | -0.70 | 2627.036 | 0.4482683 | 0.04062465 | -0.6997040 |  |
|  | -0.75 | 2607.078 | 0.4510527 | 0.04008937 | -0.7423094 |  |
|  | -0.80 | 2606.332 | 0.4169320 | 0.04242890 | -0.7978334 |  |
|  | -0.90 | 2559.172 | 0.4000274 | 0.04293319 | -0.9035296 |  |
|  | -1.00 | 2643.947 | 0.3931200 | 0.00897330 | -1.0000000 |  |

Table 1: Estimated association parameters of bivariate simulated data under different predetermined correlation coefficient at 50% censoring

The bivariate data simulated using predetermined correlation coefficient of 1.00 for instance, has maximum likelihood of 1051.85 and it produced an estimate of 0.9999. At predetermined correlation coefficients of 0.5, 0.0, -0.5 and -1.00; the maximum likelihood and the estimate of the association were (2230.212 and 0.7108), (2442.966 and 0.5443), (2528.321 and 0.4931) and (2643.947 and 0.39312) respectively. In column (e), we presented the theoretical rho (ρ) using the model but without any censoring in an attempt to check if the model runs well. Under normal conditions, Column (e) is expected to be the same with column (a). The standard errors of the estimated rhos were lower at the tails of the predefined correlation coefficients than at the centre.

With 0% censoring simulated bivariate data, the maximum likelihood of -2347.113 produced an estimate (rhos) of 0.9999 for the association parameter. At predetermined correlation coefficients of 0.5, 0.0, -0.5 and 1.00, the maximum likelihoods and the estimated association parameters were (2683.353 and 0.5160), (2836.526 and -0.0068), (2591.212 -0.5047) and (2670.088 and -0.9999) respectively as shown in Table 2Figure and Figure 2 .

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Simulation | Likelihood of | Estimated rho (ρ) | Standard Error | Theoretical |  |
|  | Rho (ρ) | function(rho) | at 0% Censoring | se(ρ) | Rho (ρ) |  |
|  | +1.00 | -2347.113 | 0.9999900 | 0.00099733 | 1.0000000 |  |
|  | +0.90 | 2002.629 | 0.9002412 | 0.00873356 | 0.8996096 |  |
|  | +0.80 | 2327.666 | 0.7971760 | 0.01768725 | 0.7955383 |  |
|  | +0.75 | 2407.804 | 0.7547523 | 0.02133885 | 0.7526036 |  |
|  | +0.70 | 2556.406 | 0.6933334 | 0.02603133 | 0.6945606 |  |
|  | +0.60 | 2581.646 | 0.5971366 | 0.03483548 | 0.6033463 |  |
|  | +0.50 | 2683.353 | 0.5160206 | 0.04040611 | 0.4931400 |  |
|  | +0.40 | 2709.906 | 0.3930111 | 0.05043064 | 0.3954632 |  |
|  | +0.30 | 2855.411 | 0.2573943 | 0.05366546 | 0.3052653 |  |
|  | +0.25 | 2860.015 | 0.2259077 | 0.05546462 | 0.2512262 |  |
|  | +0.20 | 2829.732 | 0.1659318 | 0.05913634 | 0.2151449 |  |
|  | +0.10 | 2825.826 | 0.0779796 | 0.06197048 | 0.1084885 |  |
|  | 0.00 | 2836.526 | -0.0068995 | 0.06089646 | 0.0039800 |  |
|  | -0.10 | 2792.753 | -0.1189781 | 0.06304688 | -0.1118956 |  |
|  | -0.20 | 2863.845 | -0.2082506 | 0.05564322 | -0.2157024 |  |
|  | -0.25 | 2858.524 | -0.2549800 | 0.05440402 | -0.2514401 |  |
|  | -0.30 | 2781.511 | -0.3154226 | 0.05340402 | -0.3169049 |  |
|  | -0.40 | 2773.285 | -0.3976790 | 0.04774560 | -0.4039177 |  |
|  | -0.50 | 2670.088 | -0.5047755 | 0.04174986 | -0.4941700 |  |
|  | -0.60 | 2591.212 | -0.6244719 | 0.03206368 | -0.6145968 |  |
|  | -0.70 | 2521.712 | -0.6844772 | 0.02718441 | -0.6997040 |  |
|  | -0.75 | 2392.502 | -0.7518147 | 0.02171343 | -0.7423094 |  |
|  | -0.80 | 2339.609 | -0.7964202 | 0.01771148 | -0.7978334 |  |
|  | -0.90 | 1996.704 | -0.9030565 | 0.00848095 | -0.9035296 |  |
|  | -1.00 | -2378.517 | -0.9999600 | 0.00873356 | -1.0000000 |  |

Table 2: Maximum likelihood estimates of association parameters of bivariate simulated data under different predetermined correlation coefficient at 0% censoring

Figure 1: The likelihood of the estimates of the association parameters of bivariate simulated data at predetermined correlation coefficient at 0% and 50% censoring

Figure 2: The maximum likelihood estimates of the association parameters of bivariate simulated data at predetermined correlation coefficient at 0% and 50% censoring

Discussions Using bivariate simulated data at sets of pre-determined correlation coefficients, we monitored the process and validated the model earlier developed [18]. We found that the model responded to the dictates of the inherent censoring structures and the nature of dependency between the paired set of observations. Under 50% censoring, the maximum likelihood estimates of the association parameter tended towards zero as the predetermined correlation coefficients fell from +1.00 to -1.00 but at 0% censoring the estimates were approximately the same with their respective simulation correlation coefficients. The earlier study [18]had developed a model for measuring association between bivariate censored outcomes where they had used the measurement of association between recurrence times of infections of kidney among human subjects in North East England as a case study. They found that association existed between the censored pairs of recurrence times of kidney infection and it was maximized at 0.2676.

In an attempt to check the procedure of the model and also to validate the model[18], the present study used the simulation approach to obtain standardized bivariate data under a set of predetermined correlation coefficients and various censoring conditions. We found that in the n=1000 simulated bivariate data randomly indexed as either observed or censored with 50%chances revealed an interesting trend about the estimated correlation coefficient of the n=1000 simulated bivariate censored data; as the predetermined correlation coefficients fell from +1.00 through 0.00 to -1.00, the maximum likelihood estimates of the association parameter of the bivariate censored data tended to zero but didn’t go below zero.

Our findings were different when all the n=1000 simulated bivariate data were subjected to 0% censoring. That is, when all data were presumed to be observed (no censoring), the maximum likelihood estimates were approximate of the predetermined correlation coefficients Table 2. These finding suggested that the model worked well beside been robust. In the absence of censoring, one would expect the same (or at least an approximate) correlation coefficient between the paired data irrespective of the method used in estimating the association parameter. So, the fact that the model returned an approximate estimate as the simulation predetermined correlation coefficient indicated a good performance of the model.

It was also observed during the check of the procedure that positive maximum likelihood estimates were returned even when the predetermined correlation coefficients were negative in the case of 50% censoring. One possibility for this is that each variable were subjected to 50% censoring- that is each variable were given equal chances of been censored. Although 50% censoring is a valid censoring chance, the real life situation might be different. The maximum likelihood estimates could have been negative if the proportion of the censored data were quite higher than the proportion of observed data used for the check of the model procedure.

In the same pattern, all the estimates were positive at 50% censoring, even at (ρ=-1.0). This could be attributed to the fact that a pair of bivariate data is mostly obtained from a single individual or related individuals, which suggest a form of dependency. For example, a study of association between the periods of time newly born twins will live before been infected with a particular disease requires a follow-up study of twins who are naturally related. So there is high chance of dependency between the paired observation times. Furthermore, previous studies in similar settings produced similar results. In a study of characterization of the correlation between ages at entry into Breast and Pubic Hair development[15]using interval censoring approach for measuring correlations in the presence of interval censored data[14], ascertained that the likelihood was maximized at correlation (ρ) = 0.503 to 0.506.

The model developed in [18]gave a satisfactory result under the two different censoring conditions. The model will work in similar settings in both clinical and public health sciences as well as other fields of study where the occurrence of bivariate censored outcomes is imminent.

References

1. Alejandro QF. Copula Functions and Bivariate Distributions for Survival Analysis: An Application to Political Survival. Alejandro, Q. F., (2008), Copula Functions and Bivariate Distributions for Survival Analysis: An Application to Political Survival pg 6-7. 2008.

2. Dabrowska DM, Zhang Z, Duy DL. Graphical Comparisons of Bivariate Survival Functions With an Application to Twin Studies. Ann l’institut Henri Poincar Probab Stat. 1995;31(3):545–97.

3. Dabrowska DM, Doksum K, J. S. Graphical comparison of cumulative hazards for two populations. Biometrika. 1989;76(4):763–73.

4. Pruitt RC. On negative mass assigned by the bivariate Kaplan Meier esti-mator. Ann Stat. 1991;19:443–553.

5. Prentice RL, Cai J. Covariance and survival function estimation using censored multivariate failure time data. Biometrika. 1992;79(1):495–512.

6. Dabrowska DM. Kaplan-Meier estimate on the plane. Ann Stat. 1988;16:1475–89.

7. Trivedi PKZ, David M. Copula modeling: An introduction for practitioners. Found Trends Econom. 2005;1(1):1–6.

8. Hougaard P, Harvald B, Holm N V. Assessment of dependence in the life times of twins. In Survival Analysis: State of the Art. Kluwer Academic Publisher; 1992.

9. Wang W, Adam Ding A. Estimation of Kendalls tau under censoring. Stat Sin. 2000;10:1199–215.

10. Brown BW, Hollander M, Korwar RM. Nonparametric tests of independence for censored data, with applications to heart transplant studies. Reliab Biometry. 1974;13:33–6.

11. Weier DR, Basu AP. An investigation of Kendalls t modified for censored data with applications. J Stat Plann Inference. 1980;4:381–90.

12. Oakes D. A concordance test for independence in the presence of censoring. Biometrics. 1982;38:451–5.

13. Von Mises R. On the asymptotic distribution of diﬀerentiable statistical func-tions. Ann Math Stat. 1947;18:309–48.

14. Wang W, Adam Ding A. On assessing the association for bivariate current status data. Biometrika. 2000;87(4):879–93.

15. Christensen KY, Maisonet M, Rubin C, Flanders WD, Drews-Botsch C, Dominguez C, et al. Characterization of the Correlation Between Ages at Entry Into Breast and Pubic Hair Development. Ann Epidemiol. 2010;20(5):405–8.

16. Groeneboom P, WellNer JA. Information bounds and Non-parametric Maximum Likelihood Estimation. Boston: Birkhauser; 1992.

17. Arkady S, Heekyung Y. Bayesian Estimation of Joint Survival Functions in Life Insurance. ISBA 2000. 2001. p. 1000–4.

18. Fagbamigbe AF, Adebowale AS. A model for measuring association between bivariate censored outcomes. J Mod Math Stat. 2010;4(4):127 –136.