The International Journal of Biostatistics

Volume 4, Issue 1

2008

Article 7

On the Plackett Distribution with Bivariate Censored Data

Debashis Ghosh, *Penn State University*

Recommended Citation:

Ghosh, Debashis (2008) "On the Plackett Distribution with Bivariate Censored Data," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 7.

DOI: 10.2202/1557-4679.1099

On the Plackett Distribution with Bivariate Censored Data

Debashis Ghosh

Abstract

In the analysis of dependence of bivariate correlated failure time data, a popular model is a gamma frailty model proposed by Clayton and Oakes. An alternative approach is using a Plackett distribution, whose dependence parameter has a very appealing odds ratio interpretation for dependence between the two failure times. In this article, we develop novel semiparametric estimation and inference procedures for the model. The asymptotic results of the estimator are developed; in addition, a goodness of fit test is also developed. We also discuss a regression extension to adjust for covariates using the linear regression model as well as applications to semi-competing risks data. The performance of the proposed techniques in finite samples is examined using simulation studies. Several real-data examples are used to illustrate the methodology.

KEYWORDS: accelerated failure time model, association, clustered failure times, conditional spearman rho, global cross-ratio, multivariate survival data

Author Notes: The author would like to acknowledge comments of the faculty in the Department of Biostatistics at Johns Hopkins University, where an earlier version of the material in the manuscript was presented.

1. INTRODUCTION

Clustered failure time data arise in many scientific studies. Examples include ophthalmological studies, where times to blindness in right and left eyes are correlated, family studies, where correlation is present due to shared environmental and genetic factors, and group randomized trials. Other disciplines in which such data arise include finance, engineering and political science.

We will focus on the case of bivariate censored data. Our interest is on semiparametric procedures for the analysis of such data. The dominant model that has been used is the Clayton-Oakes frailty model (Clayton, 1978; Oakes, 1982, 1986). Regression extensions of this model have been proposed by many authors, including Prentice and Hsu (1997), Glidden and Self (1999) and Fine and Jiang (2000). Research has proceeded along two lines.

Recently, there has been interest in exploring alternative dependence models. One such model is based on the Plackett distribution (Plackett, 1965). As discussed in Ghosh (2006), the dependence parameter in this model has an appealing interpretation in terms of the global cross ratio between two times to event. It has been used by many authors in the situation where there is no censoring in the data (e.g., Mardia, 1967; Dale, 1986; Molenberghs and Lesaffre, 1994; Heagerty and Zeger, 1996). With bivariate failure time data, this model has been used extensively by Burzykowski and co-authors (2001) for assessing surrogacy of one time-to-event endpoint with respect to another time-to-event endpoint.

Ghosh (2006) proposed a two-stage estimation procedure for the Plackett model in the absence of covariates, similar to that proposed by Shih and Louis (1995) and Glidden and Self (1999). This approach was also used for the same dependence model with marginal proportional hazards models by Burzykowski et al. (2001). The two-stage procedures require estimation of infinite-dimensional nuisance parameters. From a computational point of view, it would be desirable to have procedures that avoid this estimation. Such an approach would be advantageous in problems where it is not straightforward to estimate the relevant marginal distributions. This occurs in the problem of semi-competing risks (Fine et al., 2001); other scenarios are given in Martin and Betensky (2005). In this article, we develop semiparametric estimation and inference procedures for estimating the association

parameter in the Plackett distribution with bivariate censored survival data. An estimation procedure is provided that does not require estimation of the marginal distributions and allows for arbitrary dependence between censoring times. The structure of this paper is as follows. In Section 2, we define the data structures and describe the bivariate Plackett distribution. The proposed estimation procedure and associated asymptotic results are then given in Section 3. A goodness of fit test for the model is developed in Section 4. An extension to regression settings to incorporate covariates is provided in Section 5. The estimation procedure has potential advantages in the analysis of semi-competing risks data (Fine et al., 2001); we discuss this problem in Section 6. The finite-sample properties of the proposed methodologies are assessed using simulation studies; we report on their results in Section 7. Section 8 demonstrates application of the proposed methods to several datasets utilized in the literature. We conclude with some discussion in Section 9.

2. DATA STRUCTURES AND MODEL

Let (S, T) denote the times for the two events of interest, and let (A, B) be the bivariate censoring times. Define $u \wedge v$ as the minimum of u and v . Because of censoring, we observe $(X_i, Y_i, \delta_i^X, \delta_i^Y)$, $i = 1, \dots, n$, a random sample from $(X, Y, \delta^X, \delta^Y)$, where $X = S \wedge A$, $Y = T \wedge B$, $\delta^X = I(S \leq A)$, and $\delta^Y = I(T \leq B)$. We assume that (S, T) is independent of (A, B) .

The Plackett family of distributions (Plackett, 1965) is also referred to as constant global cross ratio distributions, or contingency-type distributions. Several authors have discussed estimation procedures in this model in the absence of censoring (e.g., Mardia, 1967; Dale, 1986; Molenberghs and Lesaffre, 1994; Heagerty and Zeger, 1996; Burzykowski et al., 2001). The Plackett distribution can be equivalently expressed as a copula model (Nelsen, 1999). Let $\bar{H}(x, y) \equiv Pr(S > x, T > y)$ denote the joint survivor function for (S, T) . Then based on the Plackett model, $\bar{H}(x, y) = C_\theta\{\bar{F}(s), \bar{G}(t)\}$, where for

$$u, v \in [0, 1] \times [0, 1],$$

$$C_\theta(u, v) = \begin{cases} \frac{\{1 + (\theta - 1)(u + v)\} \pm [\{1 + (\theta - 1)(u + v)\}^2 - 4uv\theta(\theta - 1)]^{1/2}}{2(\theta - 1)}, & \theta > 0, \theta \neq 1 \\ uv, & \theta = 1, \end{cases} \quad (1)$$

$\bar{F}$ is the survivor function corresponding to the marginal distribution of S , and $\bar{G}$ is the survivor function corresponding to the marginal distribution of T . Note that from (1), the case $\theta = 1$ corresponds to independence. If $0 < \theta < 1$, then S and T are negatively associated, while if $\theta > 1$, S and T are positively associated.

This is a semiparametric model in that $\bar{F}$ and $\bar{G}$ are infinite-dimensional parameters, while θ represents the parametric component. While Ghosh (2006) developed an estimation methodology using the two-stage approach of Shih and Louis (1995), we seek to develop a procedure that avoids estimation of $\bar{F}$ and $\bar{G}$.

For a detailed derivation of the model, along with a characterization of its properties, the interested reader is referred to Ghosh (2006). This model was proposed by Clayton (1978) as alternative to the gamma frailty model but was described as being harder to study. In addition, the range of dependence of the Plackett distribution is unconstrained, as it attains the so-called Fréchet bounds:

$$u \wedge v \leq C_\theta(u, v) \leq \max(u + v - 1, 0). \quad (2)$$

One obtains the copula on the left-side of (2) by letting θ go to infinity, while the right-hand side obtains by letting θ go to zero.

In terms of dependence, the parameter θ has an interpretation as a global cross-ratio parameter; this is to be contrasted with the local cross-ratio parameter in the Clayton-Oakes model. The global cross-ratio function, evaluated at (s, t) is given by

$$\begin{aligned} \theta(s, t) &= \frac{Pr(S > s | T > t) Pr(S \leq s | T \leq t)}{Pr(S > s | T \leq t) Pr(S \leq s | T > t)} \\ &= \frac{Pr(S > s, T > t) Pr(S \leq s, T \leq t)}{Pr(S > s, T \leq t) Pr(S \leq s, T > t)} \end{aligned}$$

The interpretation of the global odds ratio is an extension of the ordinary odds ratio in logistic regression. The global odds ratio is the odds ratio of having event 1 by time s given event 2 occurs by time t ; the Plackett model assumes this quantity to be constant.

It is useful to contrast this model with the Clayton-Oakes model. First, the Clayton-Oakes model assumes the constancy of a much different parameter which we call the local cross-ratio function (Oakes, 1989). In some situations, the odds ratio interpretation of the global cross-ratio function is potentially more appealing. The other major advantage of the Plackett model is that it allows for much greater negative dependence than the Clayton-Oakes model. In the notation of equation (2), the Clayton-Oakes model can achieve upper Fréchet bound, but it does not exist as a proper joint distribution function at the lower Fréchet bound. By contrast, the Plackett model is well-defined at both the lower and upper bounds in (2).

3. ESTIMATION AND INFERENCE PROCEDURES

For sake of comparison, let us consider the Clayton-Oakes model for bivariate censored data. Estimation procedures were proposed by Clayton (1978) and Oakes (1982, 1986) using an estimating equation approach based on Kendall's tau. The definition of this quantity is

$$\tau = P\{(S_1 - S_2)(T_1 - T_2) > 0\} - P\{(S_1 - S_2)(T_1 - T_2) < 0\}.$$

For the Clayton-Oakes model, this quantity is easily calculable. However, no such closed form solution exists for the Plackett model. Instead, we will construct an estimation procedure based on Spearman's ρ . The definition of Spearman's rho is given by

$$\rho \equiv 3[P\{(S_1 - S_2)(T_1 - T_3) > 0\} - P\{(S_1 - S_2)(T_1 - T_3) < 0\}]. \quad (3)$$

We first assume that there is no censoring, or equivalently, that $P(S \leq A) = P(T \leq B) = 1$. Since we have assumed $(S_1, T_1), \dots, (S_n, T_n)$ are to be a random sample from a continuous bivariate distribution, then $\rho = 6P\{(S_1 - S_2)(T_1 - T_3) > 0\} - 3$. In terms of a general copula function $C(u, v)$,

ρ can be expressed as

$$\begin{aligned}\rho &= 12 \int_0^1 \int_0^1 uv dC(u, v) - 3 \\ &= 12 \int_0^1 \int_0^1 C(u, v) dudv - 3.\end{aligned}\quad (4)$$

By plugging in (1) into (4), it can be shown that

$$\rho \equiv g(\theta) = \frac{\theta + 1}{\theta - 1} - \frac{2\theta}{(\theta - 1)^2} \ln \theta. \quad (5)$$

for the Plackett family of copulas. Equations (3) and (5) imply that

$$P\{(S_1 - S_2)(T_1 - T_3) > 0\} = \frac{g(\theta) + 3}{6}. \quad (6)$$

This motivates

$$\sum_{i < j < k} [I\{(S_i - S_j)(T_i - T_k) > 0\} - h(\theta)],$$

where $h(\theta) = \{g(\theta) + 3\}/6$, as an estimating function for θ in the Plackett model when there is no censoring.

Because of censoring, $I\{(S_i - S_j)(T_i - T_k) > 0\}$ is not necessarily observable for censored subjects. Our approach is to use instead a local version of Spearman's rho and to consider only comparable triplets of bivariate observations. Define $S_{ij} = S_i \wedge S_j$ and $T_{ik} = T_i \wedge T_k$. We let $\text{sgn}(a)$ be $-1, 0$ and 1 if a is negative, zero and positive, respectively. We take

$$\rho_l(s, t) \equiv 3E\{\text{sgn}(S_i - S_j)(T_i - T_k) | S_{ij} = s, T_{ik} = t\}. \quad (7)$$

as a local version of Spearman's rho. Given that (1) holds, $\rho_l(s, t)$ has the same value as in (5). We next define $(S_i, T_i), (S_j, T_j), (S_k, T_k)$ to be comparable if $S_{ij} \leq A_i, S_{ij} \leq A_j, T_{ik} \leq B_i$ and $T_{ik} \leq B_k$, where $X_{ij} = X_i \wedge X_j, Y_{ik} = Y_i \wedge Y_k, \{i, j, k\} \in \{1, \dots, n\}$. Let D_{ijk} denote this event. Then it follows that $E\{[I\{(S_i - S_j)(T_i - T_k) > 0\} - h(\theta)] | D_{ijk}\} = 0$. Since $S_{ij} = X_{ij}$

and $T_{ik} = Y_{ik}$ when D_{ijk} is true, we have by conditional expectation arguments,

$$\begin{aligned} E \{ I(D_{ijk} = 1) [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta)] \} &= \\ E [E \{ I(D_{ijk} = 1) [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta)] \} | D_{ijk}] &= \\ = E \{ I(D_{ijk} = 1) E [I\{(S_i - S_j)(T_i - T_k) > 0\} - h(\theta)] | D_{ijk} \} &= 0. \end{aligned}$$

This motivates the following class of estimating functions

$$U(\theta) = \sum_{i < j < k} W_{ijk} I(D_{ijk} = 1) [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta)], \quad (8)$$

where $W_{ijk} \equiv W(X_i, X_j, X_k, Y_i, Y_j, Y_k)$,

$$h(\theta) = \{g(\theta) + 3\}/6,$$

and $W(a, b, c, d, e, f)$ is a bounded weight function that converges uniformly to a deterministic and bounded weight function $w(a, b, c, d, e, f)$. Note that W can be a data-dependent function. Define $\hat{\theta}$ to be solution from setting $U(\theta) = 0$.

To derive the limiting distribution of $n^{1/2}(\hat{\theta} - \theta_0)$, we first must prove consistency of $\hat{\theta}$ for θ_0 . We know that in a neighborhood of θ_0 , $n^{-3}U(\theta) - n^{-3}u(\theta)$ converges uniformly to zero, where

$$u(\theta) = \sum_{i < j < k} w_{ijk} I(D_{ijk} = 1) [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta)].$$

Thus, the limit of θ^* , the solution of $u(\theta) = 0$, is the same as that of $\hat{\theta}$. Under model (5), $E[u(\theta_0)] = 0$. By the strong law of large numbers for U-statistics (Van der Vaart 2000, p. 192) and the continuous mapping theorem, $\hat{\theta}$ is consistent for θ_0 . In the Appendix, we prove the following theorem.

Theorem 1: (a) Under the usual regularity conditions, $n^{-5/2}U(\theta_0)$ converges in distribution to a normal random variable with mean zero and variance

$$J \equiv J(\theta) = 9 \text{ Cov } \{f(X_1, Y_1, X_2, Y_2, X_3, Y_3), f(X_1, Y_1, X_4, Y_4, X_5, Y_5)\},$$

where $f(X_i, Y_i, X_j, Y_j, X_k, Y_k) = w_{ijk} I(D_{ijk} = 1) I\{(X_i - X_j)(Y_i - Y_k) > 0 - h(\theta_0)\}$.

(b) $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to a normal random variable with mean zero and variance $I^{-2}J \equiv I(\theta_0)^{-2}J(\theta_0)$, where

$$I(\theta) = \lim_{n \rightarrow \infty} -n^{-3} \sum_{i < j < k} w_{ijk} I(D_{ijk} = 1) h'(\theta),$$

and $h'(\theta)$ is the derivative of $h(\theta)$.

In principle, consistent estimators of I and J , $\hat{I}$ and $\hat{J}$, can be constructed using sample-based quantities. The formulas for $\hat{I}$ and $\hat{J}$ can be found in the Appendix. In addition, we can construct a level α score test for the Plackett model by calculating $\{U(\theta_0)/\hat{J}(\theta_0)\}^2$ and comparing its value to the $(1 - \alpha)$ th percentile of a χ_1^2 distribution.

Because the estimating function for θ_0 is based on a U-statistic whose kernel is of order 3, the estimator of J will be computationally intensive. As an alternative, we use the bootstrap in order to estimate the variance of $\hat{\theta}$ and to construct confidence intervals for θ_0 . The finite-sample properties of the proposed method are studied in Section 7.

4. GOODNESS OF FIT TEST

In this section, we propose a test for assessing the adequacy of the Plackett model for bivariate censored data. The idea is that if the model (5) holds, then different choices of weight functions in the estimating function (12) will lead to consistent estimators of θ . Similar approaches have been taken in other contexts by Lin (1991) and Shih (1998).

Let us next consider $\hat{\theta}_1$ and $\hat{\theta}_2$ denote the solution of $U(\theta) = 0$ with weights W_1 and W_2 in (10). Under the null hypothesis that the Plackett model holds, we prove in the Appendix that $n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2)$ converges in distribution to a normal random variable with variance V . The formula for V is in the Appendix. A consistent estimator for V , say $\hat{V}$, can be found by plugging in sample-based quantities; we provide a formula for it in the Appendix. We will again utilize the bootstrap in order to construct a null distribution for $n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2)$. We generate B datasets resampling with replacement from $\{(X_1, \delta_1^X, Y_1, \delta_1^Y), \dots, (X_n, \delta_n^X, Y_n, \delta_n^Y)\}$. We can then solve the estimating function (12) for the b th dataset ($b = 1, \dots, B$) with weights W_1 and W_2 to give estimates θ_{b1}^* and θ_{b2}^* . We can then get a p-value for model adequacy using $B^{-1} \sum_{b=1}^B I(|\theta_{b1}^* - \theta_{b2}^*| \geq |\theta_1 - \theta_2|)$.

5. REGRESSION EXTENSION

In many settings, it is important to adjust for covariates, say a p -dimensional vector Z . The observed data structure from §2 is then augmented by covariates so that we have a random sample of size n from $(X, Y, \delta^X, \delta^Y, Z)$. Two-stage regression estimation procedures have been proposed by Hsu and Prentice (1997) and Glidden and Self (1999); these require estimation of infinite-dimensional parameters, which we wish to avoid. This is in keeping with the estimation procedure developed in §3.

We adopt the approach of Fine and Jiang (2000) and postulate marginal accelerated failure time models for S and T :

$$\log S = \alpha'Z + \epsilon_1 \quad (9)$$

and

$$\log T = \beta'Z + \epsilon_2 \quad (10)$$

where α and β are p -dimensional vectors of unknown regression coefficients to be estimated, and (ϵ_1, ϵ_2) is a bivariate vector of error terms with mean zero, marginal distribution functions F_1 and F_2 and copula $C_\theta(u, v)$ from (1).

Fine and Jiang (2000) showed that $\log S - \alpha'Z$ and $\log T - \beta'Z$ do not depend on $C_\theta(u, v)$. This leads to the following two-stage estimation procedure. First, we estimate α and β in (9) and (10) using the data (X_i, δ_i^X, Z_i) and (Y_i, δ_i^Y, Z_i) , $i = 1, \dots, n$, respectively. Next, we calculate residuals from the fitted regression models and apply the estimation procedure of §3 to them.

To estimate the regression coefficients, we will utilize the procedure of Jin et al. (2003). The regression parameters are estimated by the following algorithm; we illustrate the algorithm for (9):

1. Compute an initial estimate of α using the Gehan-based rank estimating function:

$$\tilde{U}(\alpha) = \sum_{i=1}^n \delta_i^X S^{(0)}\{\alpha; e_i(\alpha)\} \left[Z_i - \frac{S^{(1)}\{\alpha; e_i(\alpha)\}}{S^{(0)}\{\alpha; e_i(\alpha)\}} \right],$$

where $e_i(\alpha) = \log X_i - \alpha'Z_i$ and $S^{(j)}(\alpha, t) = n^{-1} \sum_{i=1}^n I\{e_i(\alpha) \geq t\} X_i^j$, $j = 0, 1$. Let $\tilde{\alpha}$ denote the solution from setting $\tilde{U}(\alpha)$ equal to zero.

2. Based on the estimate $\tilde{\alpha}$, construct a new estimating function

$$\tilde{U}_2(\alpha; \tilde{\alpha}) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n W_i(\tilde{\alpha}) \delta_i^X I\{e_i(\alpha) \leq e_j(\alpha)\},$$

where W_i is a data-dependent weight function that is free of α . Solve $\tilde{U}_2$ for zero to get a new estimator $\tilde{\alpha}_2$.

3. Plug $\tilde{\alpha}_2$ for $\tilde{\alpha}$ into $\tilde{U}_2$ to get a new estimator $\tilde{\alpha}_3$. Keep repeating this step until convergence. Denote the estimator at convergence by $\hat{\alpha}$.

The algorithm would work similarly for β . In Jin et al. (2003), they derived the consistency and asymptotic normality of $\hat{\alpha}$. The variance of the limiting distribution of $n^{1/2}(\hat{\alpha} - \alpha)$ is difficult to estimate directly, since it involves the derivative of the density of the error term ϵ_1 . A resampling methodology is proposed in Jin et al. (2003) to construct confidence intervals; we utilize that same approach here as well. Because the inference for the regression coefficients and dependence parameter are completely orthogonal in this model, we can directly apply Theorem 1 here to perform inference on θ .

6. APPLICATION TO SEMI-COMPETING RISKS

Even though the proposed methods are computationally simple and do not require estimation of infinite-dimensional nuisance parameters, they will in general be less efficient than the two-stage approach. In this section, we give an example of a particular data structure in which it is advantageous to estimate the dependence parameter using a modification of the methods in §3. This is referred to as semi-competing risks data (Fine et al., 2001).

In this situation, we observe the data $(R_i, \delta_i^R, Y_i, \delta_i^Y)$ ($i = 1, \dots, n$), n iid copies from $(R, \delta^R, Y, \delta^Y)$, where $R = S \wedge (T \wedge C)$ and $\delta^R = I(S \leq T \wedge C)$. Note that in this situation, S is censored by the minimum of T and C . While the censoring mechanism is the same for T as in §2, it is quite different for S . This is known as “semi-competing risks” data because while T can potentially censor S , the reverse cannot happen.

Several facts obtain about this type of data structure. First, what is identifiable from the observed data is the $\bar{H}$ on the region where $0 \leq x \leq$

y . This can be estimated using the nonparametric estimator of Lin and Ying (1993): $\hat{H}(x, y) = n^{-1} \sum_{i=1}^n I(R_i > x, Y_i \geq y) / \hat{S}_C(y)$, where $\hat{S}_C$ is the Kaplan-Meier estimator of the survival function for C . To adapt the Plackett model to semi-competing risks data, we assume that (1) holds on the region where $0 \leq u \leq v$. We refer to this as the Plackett model on the wedge. The second is that the marginal distribution of S is not completely identifiable given the observed data given the order restriction. Even if we assume the Plackett model on the wedge, then $C(u, 0)$ does not have an interpretation as a marginal distribution. This can only be given a marginal distribution interpretation if we formulate a model for the region where $0 \leq v \leq u$. Finally, if the T represents a time to terminal event, then S represents the failure time if the subject had not terminated. In many medical settings, this interpretation can be quite controversial.

Starting with the definition of the local Spearman's rho in (5), we now define a new region of comparability so that we can construct an unbiased estimating function for θ . We observe that for semi-competing risks data, $(S_i, T_i), (S_j, T_j)$ and (S_k, T_k) are comparable if $S_{ij} \leq T_{ij} \leq C_{ij}$ and $S_{ik} \leq T_{ik} \leq C_{ik}$. Let $\tilde{D}_{ijk}$ denote this event. Then proceeding as in §3, we get the following estimating function for estimation of θ , similar in form to (8):

$$U^s(\theta) = \sum_{i < j < k} W_{ijk}^s I(\tilde{D}_{ijk} = 1) [I\{(R_i - R_j)(Y_i - Y_k) > 0\} - h(\theta)], \quad (8)$$

where W_{ijk}^s is a data-dependent weight function that converges uniformly to a bounded and deterministic limit w^s . Let $\hat{\theta}^s$ denote the estimator from setting $U^s(\theta) = 0$. We can modify the arguments in the proofs of Theorem 1 and 2 to derive the consistency and asymptotic normality of $\hat{\theta}^s$, along with consistent estimators of the variance of the limiting distribution of $n^{1/2}(\hat{\theta}^s - \theta)$. Based on the estimator for θ , we can then proceed as in Fine et al. (2001) or Lakhal-Chaleib et al. (2006) to derive the estimator for $\bar{F}$. That will not be pursued here.

Note that this is a situation in which estimation of the dependence is made simpler by constructing an estimating function. If one were to instead attempt nonparametric maximum likelihood estimation (NPMLE) in the Plackett model on the wedge, this becomes a much more difficult on a

variety of fronts. First, characterization of the NPMLE seems hard. Second, most estimation algorithms would involve to estimate the nuisance parameters. It also appears difficult to incorporate the identifiability issue of the marginal distribution of S in such an estimation procedure. Finally, constructing a consistent estimator for the variance of the dependence parameter might potentially be complicated.

7. SIMULATION STUDIES

To assess the finite-sample properties of the proposed estimation methods, we conducted a series of simulation studies. Sample sizes $n = 100, 200$ and 500 were considered. We took $\theta = 0.5, 1$ and 2 . Data were generated from (1) using exponential random variables with rate parameter 2 . Details on the data simulation procedure can be found in Ghosh (2006).

We considered three censoring scenarios: no censoring, independent censoring using a $\text{Uniform}(0, 2)$ random variable and independent censoring using a uniform $(0, 3)$ random variable. The moderate and light censoring situations yield approximately 20% and 12% censoring for both X and Y . For each simulation setting, 500 sample datasets were generated. The unweighted estimating equation with $W = 1$ was used here. We calculated the bootstrap variance using 200 bootstrap samples for each iteration. The results from using the estimator of θ from (12) are presented in Table 1. We find that there is little bias in the parameter estimates and that the bootstrap yields reasonable standard error estimates and coverage probabilities for 95% confidence intervals. Comparing these results to those in Ghosh (2006), we find that the proposed method is 15-20% less efficient relative to the two-stage approach.

Table 1. Summary of simulation results for θ in Plackett model with proposed estimator

n	θ	Uncensored				U(0,3) censoring				U(0,2) censoring			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
100	0.5	-0.02	0.24	0.22	0.96	-0.03	0.21	0.20	0.95	-0.02	0.26	0.24	0.95
	1	-0.03	0.40	0.38	0.95	-0.02	0.37	0.32	0.95	-0.03	0.46	0.42	0.95
	2	-0.03	0.65	0.62	0.95	-0.02	0.63	0.60	0.96	-0.02	0.75	0.72	0.96
200	0.5	0.01	0.17	0.14	0.96	-0.02	0.19	0.17	0.96	-0.01	0.17	0.16	0.95
	1	-0.01	0.29	0.26	0.95	-0.02	0.28	0.26	0.95	-0.02	0.32	0.28	0.94
	2	0.01	0.47	0.44	0.96	0.01	0.51	0.50	0.96	-0.02	0.55	0.53	0.95
500	0.5	-0.01	0.12	0.10	0.96	-0.01	0.11	0.09	0.94	-0.01	0.11	0.10	0.94
	1	-0.01	0.20	0.18	0.94	0.01	0.16	0.14	0.94	0.01	0.21	0.19	0.94
	2	-0.01	0.32	0.31	0.95	0.01	0.33	0.32	0.94	0.01	0.34	0.33	0.95

Note: Bias is the mean of the estimators of θ_0 minus θ_0 ; SE is the standard error of the estimators of θ_0 ; SEE is the mean of the standard error estimate; CP is the coverage probability of the 95% confidence interval based on the bootstrap. Independent censoring was generated in two ways: using a Uniform (0, 3) random variable and a Uniform (0, 2) random variable.

In the next set of simulation studies, we studied the finite-sample properties of the proposed goodness of fit method developed in Section 4. Sample sizes $n = 200$ and $n = 500$ were considered. We again looked at the scenario of no censoring, $U(0, 2)$ and $U(0, 3)$ censoring. We assessed size using the Plackett model with $\theta = 1$ and $\theta = 2$. For studying power, we generated data from a Clayton-Oakes model with $\eta = 2$ and $\eta = 5$. We used the weight functions $W_1 = 1$ and

$$W_2 = I(X_i \geq a, Y_i \geq b, X_j \geq c, Y_j \geq d, X_k \geq e, Y_k \geq f)$$

to construct estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. The summary of the simulation results is given in Tables 2 and 3. We find that the procedure achieves the proper size and is relatively powerful in the scenarios considered.

Table 2. *Empirical sizes of proposed goodness of fit method*

n	θ	Uncensored	U(0,3) censoring	U(0,2) censoring
200	1	0.047	0.049	0.052
	2	0.048	0.048	0.051
500	1	0.050	0.051	0.049
	2	0.048	0.048	0.049

Note: Data generated using Plackett model (1).

Table 3. *Empirical powers of proposed goodness of fit method*

n	η	Uncensored	U(0,3) censoring	U(0,2) censoring
200	2	0.30	0.28	0.25
	5	0.70	0.65	0.60
500	2	0.52	0.49	0.48
	5	0.82	0.79	0.77

Note: Data generated using Clayton model.

8. NUMERICAL EXAMPLES

We now present some applications of the methodologies described in the article. The estimate of the standard errors for θ was based on the bootstrap distribution of $\hat{\theta}$ using 1000 simulated datasets. Similarly, the p-value for the goodness of fit test from §4 was based on the bootstrap distribution of the test statistic applied to 1000 simulated datasets.

8.1. Diabetic Retinopathy Data

The first dataset we analyze is the diabetic retinopathy dataset considered by Huster et al. (1989) among others. The goal of the study was to see if laser photocoagulation treatment was effective in delaying the onset of blindness in patients with diabetic retinopathy. For each patient, treatment was randomly assigned to one eye; for the other eye, no treatment was performed. The time to event for each eye was first occurrence of visual acuity less than 5/200. There are data on 197 patients; a covariate that was measured was type of diabetes (adult versus juvenile).

The plot of the time to event based on type of diabetes, stratified by treatment, is given in Figure 1.

Based on the plots, we find that the untreated group (Figure 1(b)) tends to have increased risk of the event relative to the treated group. Second, the effect of type of diabetes is different for the treated group compared to the untreated group. For the treated eyes, the early onset eyes have higher risk of blindness, but this is reversed for the untreated eyes. It should be noted that the difference is not statistically significant at a level of 0.05 (log-rank p-value corresponding to Figure 1(a) = 0.08; log-rank p-value corresponding to Figure 1(b) = 0.08).

If we ignore the type of diabetes, then the estimate of θ from the Plackett model is 4.9. The associated standard error is 0.20. The resulting Wald statistic indicates strong positive dependence between the time to events in the two eyes. We performed a goodness of fit test using the method described in §4. The weighted estimating function yield a value of θ of 4.74 with an associated standard error of 0.15; the p-value associated with the test statistic from §4 is 0.34. Thus, we have no significant evidence of lack of fit for the Plackett model.

Next, we performed an analysis in which we adjusted for the type of diabetes. We took Z to be a binary indicator with $Z = 0$ representing early onset diabetes and $Z = 1$ indicating late onset diabetes. We let S denote time to blindness for treated eyes and T denote time to blindness for untreated eyes. Using the approach of Jin et al. (2003), we obtain $\hat{\alpha} = 0.86$, with an associated standard error of 0.42, while the corresponding values for β are -0.39 and 0.27 . This implies that relative to people with early onset

Figure 1: Kaplan-Meier curves of time to occurrence of visual acuity less than 5/200 for the treated (Figure 1a) and untreated (Figure 1b) groups based on the diabetic retinopathy data. For each plot, the solid line indicates the early onset (juvenile diabetes) group, while the dashed line indicates the late onset (adult diabetes) group.

diabetes, those with late onset diabetes have longer times to blindness in the treated group and shorter times to blindness in the untreated group.

After adjusting for the effect of diabetes, the estimate of θ is 3.52, which suggests that there is some reduction in the positive dependence between the age at onset between the two eyes if we adjust for diabetes type. The associated standard error is 0.13, still implying a statistically significant positive association. The weighted estimate of θ is 3.5 with a standard error of 0.11; application of the method in §4 yields a p-value of 0.53. This implies that even after adjusting for covariates, there is no evidence against the Plackett model.

8.2. Transplantation Data

In this section, we consider data from a transplantation study (Copelan et al., 1991). In this multicenter clinical trial, patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) underwent bone marrow transplantation and were followed prospectively. As noted in Klein and Moeschberger (1997, Section 1.3), the recovery from transplantation is quite complex. In the ideal scenario, then the patient recovers without any side effects, and the leukemia enters remission. However, if the body rejects the transplant, then acute graft versus host disease may develop, which leads to increased risk of death or relapse. These data have been analyzed by Fine et al. (2001) and by Wang (2003) in terms of assessing the dependence between time to relapse and survival for the entire population using the gamma frailty approach.

Applying the procedure in §6, we obtain $\hat{\theta} = 17.81$ with a standard error of 0.03. This implies a highly positive association between time to relapse with time to death. A modification of the goodness of fit test in Section 4, based on the estimating function in §6, $W_1 = 1$ and from Fine et al. (2001, §2)

$$W_2 \equiv W_{a,b}(s, t) = \{n^{-1} \sum_{i=1}^n I(S_i \geq a \wedge s, T_i \geq t \wedge b)\}^{-1},$$

where a and b are the 95th percentile values of X and Y , yielded a p-value of 0.54, based on 1000 bootstrapped datasets. This suggests that the Plackett model is an adequate model to fit here.

9. DISCUSSION

In this paper, we have formulated novel semiparametric estimation and inference procedures for the Plackett with bivariate survival data. Several generalizations of the methodology were given. In the bivariate censored data case, the limited simulation studies show that the methodology is less efficient than the two-stage approach of Shih and Louis (1995). It should be possible to increase efficiency using a different weight function, but that issue was not explored here and remains an open problem for further research. Finding the appropriate weight function will be difficult since it is based on the true underlying data-generating mechanism.

The major advantage we envision of the proposed estimating function approach is in situations in which the observable region is complicated. One example arises in the semi-competing risks setup, described in §6. Other situations involving complex censoring and truncation patterns are given by Martin and Betensky (2005). For these problems, estimating marginal distributions of S and T will be complicated, and approaches that avoid the estimation of infinite-dimensional parameters, such as those presented here, represent a major advance.

Ghosh (2006) described the complications of extending the model to more than two failure times. It appears that extensive numerical computation is needed. This is clearly an area where probabilistic models based on Archimedean copulas (Oakes, 1989; Genest and Rivest, 1993) are at a decided advantage. An alternative approach is to apply an extension of the proposed procedure combined with use of a composite likelihood (Lindsay, 1988).

There has been recent interest in the recent literature on use of a local version of Kendall's τ to define unbiased estimating functions with bivariate censored data under complex sampling (Martin and Betensky, 2005; Lakhali-Chaleib et al., 2006). Since such a quantity is not easily calculable with the Plackett model, we made use of a conditional Spearman's rho. It is anticipated that such a conditioning will help to address estimation problems under complex sampling schemes.

APPENDIX

Proof of Theorem 1:

(a) By addition and subtraction,

$$\begin{aligned} n^{-5/2}U(\theta_0) &= n^{-5/2} \sum_{i < j < k} w_{ijk} \delta_{ijk} [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta_0)] \\ &\quad + n^{-5/2} \sum_{i < j < k} \delta_{ijk} (W_{ijk} - w_{ijk}) [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta_0)]. \end{aligned}$$

By assumption on W and the boundedness of the censoring distribution function, the second term is $o_P(1)$. This yields

$$n^{-5/2}U(\theta_0) = n^{-5/2} \sum_{i < j < k} \psi_{ijk} + o_P(1),$$

where $\psi_{ijk} = w_{ijk} \delta_{ijk} [I\{(X_i - X_j)(Y_i - Y_k) > 0\} - h(\theta_0)]$, $1 \leq i < j < k \leq n$. The result from part (a) of Theorem 1 now follows by the Central Limit Theorem for U -statistics (Van der Vaart 2000, p.175).

(b) Taking a Taylor series expansion of $U(\hat{\theta})$ about θ_0 and applying the consistency of $\hat{\theta}$ for θ_0 , we have

$$n^{1/2}(\hat{\theta} - \theta_0) = \left[n^{-3} \frac{dU(\theta_0)}{d\theta} \right]^{-1} \{n^{-5/2}U(\theta_0)\} + o_P(1),$$

where

$$\frac{dU}{d\theta} = - \sum_{i < j < k} w_{ijk} \delta_{ijk} h'(\theta).$$

By (a), the uniform strong law of large numbers (Pollard 1990, p. 41) and the continuous mapping theorem, we have the desired result.

An estimator of J is given by

$$\begin{aligned} \hat{J} = 9n^{-6} \sum_{i < j < k < l < m} &\left(\hat{\psi}_{ijk} \hat{\psi}_{ijl} + \hat{\psi}_{ijk} \hat{\psi}_{ijm} + \hat{\psi}_{ijk} \hat{\psi}_{ikl} \right. \\ &\quad + \hat{\psi}_{ijk} \hat{\psi}_{ikm} + \hat{\psi}_{ijk} \hat{\psi}_{ilm} + \hat{\psi}_{ijk} \hat{\psi}_{jkl} \\ &\quad + \hat{\psi}_{ijk} \hat{\psi}_{jkm} + \hat{\psi}_{ijk} \hat{\psi}_{jlm} + \hat{\psi}_{ijk} \hat{\psi}_{klm} \\ &\quad \left. + \hat{\psi}_{jkl} \hat{\psi}_{jkm} + \hat{\psi}_{jkl} \hat{\psi}_{jlm} + \hat{\psi}_{jkl} \hat{\psi}_{klm} \right), \end{aligned}$$

where

$$\hat{\psi}_{ijk} = W_{ijk}\delta_{ijk} \left[I\{(X_i - X_j)(Y_i - Y_k)\} > 0 - h(\hat{\theta}) \right].$$

Similarly, I can be estimated by $\hat{I} = -n^{-3} \sum_{i < j < k} W_{ijk}\delta_{ijk}h'(\hat{\theta})$. The consistency of $\hat{I}$ for I and $\hat{J}$ for J follow by the strong consistency of $\hat{\theta}$ for θ_0 and the uniform strong law of large numbers (Pollard 1990, p. 41).

Asymptotic normality of $n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2)$:

Under the null hypothesis that model (5) holds, both $\hat{\theta}_1$ and $\hat{\theta}_2$ are strongly consistent for θ_0 using the arguments in the paper. We also have that

$$n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2) = n^{-5/2} \sum_{i < j < k} Q_{ijk},$$

where $Q_{ijk} = I_1^{-1}\psi_{ijk}^1 - I_2^{-1}\psi_{ijk}^2$,

$$I_a = - \lim_{n \rightarrow \infty} n^{-3} \sum_{i < j < k} w_{ijk}^a \delta_{ijk} h'(\theta) / r(X_{ijk}, Y_{ijk}),$$

w_{ijk}^a is the appropriate modification of ψ_{ijk} , replacing w_{ijk} with $w_{ijk}^a = \lim_{n \rightarrow \infty} W_{ijk}^a$ and J_a is the appropriate modification of J from Theorem 1 for $a = 1, 2$. By the central limit theorem for U-statistics, this implies that under the null hypothesis that the Plackett model holds,

$$n^{1/2}(\hat{\theta}_1 - \hat{\theta}_2) \rightarrow N(0, \Sigma),$$

where

$$\begin{aligned} \Sigma = \lim_{n \rightarrow \infty} 9n^{-6} \sum_{i < j < k < l < m} & (Q_{ijk}Q_{ijl} + Q_{ijk}Q_{ijm} + Q_{ijk}Q_{ikl} \\ & + Q_{ijk}Q_{ikm} + Q_{ijk}Q_{ilm} + Q_{ijk}Q_{jkl} \\ & + Q_{ijk}Q_{jkm} + Q_{ijk}Q_{jlm} + Q_{ijk}Q_{klm} \\ & + Q_{jkl}Q_{jkm} + Q_{jkl}Q_{jlm} + Q_{jkl}Q_{klm}). \end{aligned}$$

A consistent estimator for Σ can be found by plugging in sample-based quantities and applying the strong consistency of $\hat{\theta}_1$ and $\hat{\theta}_2$ and the uniform strong law of large numbers (Pollard 1990, p. 41).

REFERENCES

- BURZYKOWSKI, T., MOLENBERGHS, G., BUYSE, M., GEYS, H. & RENARD, D. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics* **50**(4), 405 – 422.
- CLAYTON, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**(1), 141 – 151.
- DALE, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**(4), 909–917.
- FINE, J. P. & JIANG, H. (2000). On association in a copula with time transformations. *Biometrika* **87**(3), 559 – 571.
- FINE, J. P., JIANG, H. & CHAPPELL, R. (2001). On semi-competing risks data. *Biometrika* **88**(4), 907 – 919.
- GENEST, C. & RIVEST, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Am. Statist. Assoc.* **88**(423), 1034 – 1043.
- GHOSH, D. (2006). Semiparametric global cross-ratio models for bivariate censored data, *Scandinavian Journal of Statistics* **33**(4), 609 – 619.
- GLIDDEN, D.V. & SELF, S. G. (1999). Semiparametric likelihood estimation in the Clayton-Oakes failure time model. *Scand. J. Statist.* **26**(3), 363 – 372.
- HEAGERTY, P. J. & ZEGER, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J. Am. Statist. Assoc.* **91**(435), 1024–1036.
- HUSTER, W. J., BROOKMEYER, R. & SELF, S. G. (1989). Modelling paired survival data with covariates. *Biometrics* **45**(1), 145–156.

- JIN, Z., LIN, D.Y., WEI, L.J. & YING, Z., Rank-based inference for the accelerated failure time model. (2003). *Biometrika* **90**(2), 341-353.
- LAKHAL-CHAIEB, L., RIVEST, L. P. & ABDOUS, B. (2006). Estimating survival under a dependent truncation. *Biometrika* **93**(3), 655 – 669.
- LIN, D. Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J. Am. Statist. Assoc.* **86**(415), 725–8.
- LIN, D. Y. & YING, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* **80**(3), 573 – 582.
- LINDSAY, B. G. (1988). Composite Likelihood Methods. *Contemporary Mathematics* **80**, 221 – 239.
- MARDIA, K. V. (1967). *Families of Bivariate Distributions*. London: Charles Griffin and Co.
- MARTIN, E. C. & BETENSKY, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *J. Am. Statist. Assoc.*, **100**(470), 484-492.
- MOLENBERGHS, G., & LESAFFRE, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *J. Am. Statist. Assoc.* **89**(426), 633-644.
- NELSEN, R. (1999). *An Introduction to Copulas*. New York: Springer.
- OAKES, D. (1982). A model for association in bivariate survival data. *J. R. Statist. Soc. B* **44**(3), 414 – 422.
- OAKES, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73**(2), 353 – 361.
- OAKES, D. (1989). Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**(406), 487 – 493.

- PLACKETT, R. L. (1965). A class of bivariate distributions. *J. Am. Statist. Assoc.* **60**(310), 516 – 522.
- POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. Hayward: IMS.
- PRENTICE, R. L. & HSU, L. (1997). Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika* **84**(2), 349 – 363.
- SHIH, J. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika* **85**(1), 189 – 200.
- SHIH, J. & LOUIS, T. A. (1995). Inference on the association parameter in copula models for bivariate survival data. *Biometrics* **51**(4), 1384 – 1399.
- VAN DER VAART, A. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- WANG, W. (2003). Estimating the association parameter for copula models under dependent censoring. *J. R. Statist. Soc. B.* **65**(1), 257-273.