Biometrika (2000), **87,** 4, *pp.* 879–893

Printed in Great Britain

On assessing the association for bivariate current status data

BY WEIJING WANG

Institute of Statistics, National Chiao-Tung University, Hsin-Chu, Taiwan, R.O.C.

wjwang@stat.nctu.edu.tw

AND A. ADAM DING

Department of Mathematics, Northeastern University, Boston, Massachusetts 02115, U.S.A.

ding@neu.edu

SUMMARY

Assuming that the two failure times of interest with bivariate current status data follow a bivariate copula model, we propose a two-stage estimation procedure to estimate the association parameter which is related to Kendall's tau. Asymptotic properties of the proposed semiparametric estimator show that, although the first-stage marginal estimators have a convergence rate of only $n^{1/3}$, the resulting parameter estimator still converges to a normal random variable with the usual $n^{1/2}$ rate. The variance of the proposed estimator can be consistently estimated. Simulation results are presented, and a community-based study of cardiovascular diseases in Taiwan provides an illustrative example.

Some key words: Copula model; Cross-sectional data; Kendall's tau; Odds ratio; Pseudolikelihood; Semiparametric estimation.

1. INTRODUCTION

1·1. *Background*

Current status data arise commonly in many studies of epidemiology, biomedicine, demography and reliability; Jewell, Malani & Vittinghoff (1994) gave two examples in AIDS studies, and Diamond & McDonald (1992) discussed examples in demography. In its univariate setting 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. Such a data structure is also called interval censoring or case I in Groeneboom & Wellner (1992, p. 35). Statistical inference methods for univariate current status data have been extensively studied. For example algorithms for the nonparametric maximum likelihood estimator of the distribution function of T were proposed by Ayer et al. (1955), Peto (1973), Turnbull (1976) and Groeneboom & Wellner (1992) under the assumption that T and C are independent. Asymptotic properties of the estimator were studied in Groenboom & Wellner (1992), who showed that it converges pointwise at rate $n^{1/3}$ to a complex limiting distribution related to Brownian motion. The efficacy of smooth functionals of the estimator was studied in Groenboom & Wellner (1992). In S. van der Geer's 1994 Ph.D. Thesis at the University of Leiden, and in Huang & Wellner (1995). Van der Laan & Robins (1998) relaxed the independent censoring assumption by incorporating covariate information which accounts for the dependence between T and C and then

proposed methods for estimating smooth functionals of the distribution function of T. Semiparametric regression methods which study the relationship between T and a set of covariates have been proposed by Finkelstein (1986), Jewell & Shiboski (1990), Klein & Spady (1993), Rabinowitz, Tsiatis & Aragon (1995), Rossini & Tsiatis (1996) and Lin, Oakes & Ying (1998), just to name a few. The papers by Huang & Wellner (1997) and Jewell & van der Laan (1997) summarise recent developments for current status data.

In this paper we consider bivariate current status data. Let (T_1, T_2) be two failure times of interest with respective marginal survival functions $S_1(.)$ and $S_2(.)$ and joint survival function $S(.,.)$. The times T_1 and T_2 are possibly correlated and their dependent relationship is of main interest. Bivariate current status occur naturally when T_1 and T_2 represent the ages of onset of two chronic diseases whose starting times are difficult to trace back precisely. In the general setting, observable data consist of the vector

$$\{C_1, C_2, \delta_1 = I(T_1 \leqslant C_1), \delta_2 = I(T_2 \leqslant C_2)\},$$

where (C_1, C_2) are random monitoring times and $I(.)$ is the indicator function. Note that, if T_1 and T_2 are measured from the same individuals, then $C_1 = C_2 = C$, which is likely to be the case in most applications. Hence in the following analysis we assume that the observed data have the form

$$\{C, \delta_1 = I(T_1 \leqslant C), \delta_2 = I(T_2 \leqslant C)\}$$

and that (T_1, T_2) are independent of C.

Our main goal is to propose inference methods for studying the relationship between T_1 and T_2. A community-based study of cardiovascular diseases in Taiwan provides an illustrative example. The study was designed to determine risk factors for cardiovascular diseases in two towns, Chu-Dung and Pu-Tze, in Taiwan. Our scientific goal is to investigate whether or not the ages of hypertension, diabetes mellitus and hypercholesterolaemia are correlated with each other. Their mutual associations may reveal important information about the underlying mechanism of cardiovascular diseases.

In § 1·2 we define notation and introduce some fundamental concepts. In § 2, we derive the proposed estimators and their asymptotic properties. We examine finite sample performance of the proposed estimators in § 3 by simulation. In § 4 the proposed methodology is applied to the aforementioned dataset. In § 5 we discuss how to extend our methods to more general data structures.

1·2. *Some preliminaries*

Let $S(s, t) = \mathrm{pr}(T_1 > s, T_2 > t)$ be the joint survival function of (T_1, T_2) and let $S_i(.)$, for $i = 1, 2$, be their marginal survival functions, respectively. Let $G(c) = \mathrm{pr}(C \leqslant c)$ be the distribution function of C and $H(c, \delta_1, \delta_2) = \mathrm{pr}(C \leqslant c, \delta_1, \delta_2)$ be the subdistribution function of the observed vector (C, δ_1, δ_2). The density or probability functions of C and (C, δ_1, δ_2) are denoted by $g(c)$ and $h(c, \delta_1, \delta_2)$, respectively.

The dependence relationship between T_1 and T_2 can be fully characterised by their joint survival function $S(s, t)$. Nonparametric estimation of the bivariate survival function for right-censored data under independent censorship has been thoroughly studied in recent years. For a review, please refer to Wang & Wells (1997). However, for the bivariate current status data discussed here, there exist different functions of $S(T_1, T_2)$ and $G(C)$ that induce the same distribution function $H(C, \delta_1, \delta_2)$ for the observed vector. Hence $S(.,.)$ is not identifiable nonparametrically and we can only estimate $S(.,.)$ under parametric or semiparametric assumptions.

Semiparametric analysis for bivariate right-censored data has been proposed by Shih & Louis (1995) and Hsu & Prentice (1996). They assume that (T_1, T_2) follow a copula model with the joint survival function

$$S(s, t) = C_\alpha\{S_1(s), S_2(t)\}, \tag{1}$$

where $C(.,.): [0, 1]^2 \to [0, 1]$, which is itself a genuine survival function on the unit square, determines the local dependence structure and $\alpha \in R$ is a global association parameter related to Kendall's tau, denoted by τ, as follows:

$$\tau = 4 \int_0^1 \int_0^1 C_\alpha(u, v)\, du\, dv - 1. \tag{2}$$

The copula family includes many useful bivariate lifetime models and has gained considerable attention in recent years because of its modelling flexibility; see Genest & McKay (1986), Oakes (1989) and Genest & Rivest (1993).

The main result of the paper focuses on semiparametric estimation of the association parameter α for bivariate current status data when the form of association $C_\alpha(.,.)$ is specified up to α but the marginals $S_i(.)$ $(i = 1, 2)$ remain unspecified. The proposed approach involves two stages of estimation. In the first stage S_i is estimated by $\hat{S}_i$ $(i = 1, 2)$. In the second stage the proposed estimator of α, denoted by $\hat{\alpha}$, is obtained by maximising a 'pseudo' likelihood estimating equation with the true S_i being replaced by $\hat{S}_i$ $(i = 1, 2)$. Similar ideas have been used by Genest, Ghoudi & Rivest (1995) for complete data and Shih & Louis (1995) for right-censored data. The properties of $\hat{\alpha}$ depend on regularity conditions on the imposed copula model and the plugged-in estimators $\hat{S}_i$ $(i = 1, 2)$. For our problem, the nonparametric maximum likelihood estimator can be a candidate for $\hat{S}_i$ $(i = 1, 2)$. However, unlike the above two papers in which $\hat{S}_i$ has the standard $n^{1/2}$ convergence rate, the convergence rate of the nonparametric maximum likelihood estimator for current status data is only $n^{1/3}$. Therefore, the asymptotic expansions of $\hat{S}_i$ $(i = 1, 2)$ are more complex in the present setting than in the cases of complete data or right-censored data. Using the results in Groneboom & Wellner (1992) and Huang & Wellner (1995) we show that, under suitable assumptions, the proposed estimator of α converges to a normal random variable at the standard rate $n^{1/2}$ and can be asymptotically expressed as a sum of independently identically distributed terms which will be useful in variance estimation.

2. Estimation for copula models

2·1. *Notation*

Let $\{T_{1i}, T_{2i}, C_i, i = 1, \ldots, n\}$ be a random sample from (T_1, T_2, C). Under the copula assumption stated in (1), the loglikelihood function is given by

$$\begin{aligned}\log L(\alpha, S_1, S_2) = {} & \sum_{i=1}^n \log g(c_i) + \sum_{i=1}^n \delta_{1i}\delta_{2i} \log S_{11}(\alpha, c_i) + \sum_{i=1}^n (1 - \delta_{1i})\delta_{2i} \log S_{01}(\alpha, c_i) \\ & + \sum_{i=1}^n \delta_{1i}(1 - \delta_{2i}) \log S_{10}(\alpha, c_i) + \sum_{i=1}^n (1 - \delta_{1i})(1 - \delta_{2i}) \log S_{00}(\alpha, c_i),\end{aligned} \tag{3}$$

where

$$S_{11}(\alpha, c) = \mathrm{pr}(T_1 < c, T_2 < c) = 1 - S_1(c) - S_2(c) + C_\alpha\{S_1(c), S_2(c)\},$$
$$S_{01}(\alpha, c) = \mathrm{pr}(T_1 > c, T_2 < c) = S_1(c) - C_\alpha\{S_1(c), S_2(c)\},$$
$$S_{10}(\alpha, c) = \mathrm{pr}(T_1 < c, T_2 > c) = S_2(c) - C_\alpha\{S_1(c), S_2(c)\},$$
$$S_{00}(\alpha, c) = \mathrm{pr}(T_1 > c, T_2 > c) = C_\alpha\{S_1(c), S_2(c)\}.$$

When the marginals S_1 and S_2 are known, a natural estimator for the association parameter α is the maximum likelihood estimator that maximises (3). The proposed two-stage estimator of α, denoted by $\hat{\alpha}$, maximises the 'pseudo' likelihood function $\log L(\alpha, \hat{S}_1, \hat{S}_2)$, where $\hat{S}_j$ is an estimator of S_j $(j = 1, 2)$.

2·2. *The estimation procedure*

In the first stage, S_j can be estimated based on the univariate sample

$$\{(C_i, \delta_{ji}),\ i = 1, \ldots, n;\ j = 1, 2\}.$$

We shall first focus on using the nonparametric maximum likelihood estimators as our first-stage estimators. Estimation of S_j $(j = 1, 2)$ by incorporating covariate information will be discussed in § 5. Let $S_j = 1 - F_j$ and $\hat{S}_j = 1 - \hat{F}_j$. The nonparametric maximum likelihood estimator of F_j, denoted by $\hat{F}_j$, maximises the function

$$l(F_j) = \sum_{i=1}^{n} \{\delta_{ji} \log F_j(C_i) + (1 - \delta_{ji}) \log S_j(C_i)\} \quad (j = 1, 2).$$

Estimator $\hat{F}_j$ solves the self-consistency equation described in Groeneboom & Wellner (1992, pp. 66–7) and can be obtained using the greatest convex minorant algorithm (Groeneboom & Wellner, 1992, pp. 40–1; Huang & Wellner, 1997, p. 127), which is considered to be faster than the EM algorithm. Note that $\hat{F}_j$ can be represented by the max–min formula

$$\hat{F}_j(c_{(i)}) = \max_{l \leqslant i} \min_{k \geqslant i} \frac{\sum_{m=l}^{k} \delta_{(jm)}}{k - l + 1},$$

where $c_{(1)} < \ldots < c_{(n)}$ are ordered observed values of $(C_1, \ldots, C_n)$ and $\delta_{(ji)}$ $(j = 1, 2)$ are the associated indicators for $C_{(i)}$.

In the second stage, we estimate the association parameter α by plugging in the first-stage estimators $\hat{S}_1$ and $\hat{S}_2$, and then maximising the resulting 'pseudo' loglikelihood function, namely

$$\log L(\alpha, \hat{S}_1, \hat{S}_2) = n \int l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\}\, dH_n(c, \delta_1, \delta_2),$$

where

$$l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\} = \delta_1\delta_2 \log S_{11}(\alpha, c) + \delta_1(1 - \delta_2) \log S_{10}(\alpha, c) + (1 - \delta_1)\delta_2 \log S_{01}(\alpha, c) + (1 - \delta_1)(1 - \delta_2) \log S_{00}(\alpha, c),$$

and $H_n(c, \delta_1, \delta_2)$ denotes the empirical estimator of $H(c, \delta_1, \delta_2)$. Equivalently $\hat{\alpha}$ solves the score equation

$$U(\alpha, \hat{S}_1, \hat{S}_2, H_n) = \frac{1}{n}\frac{\partial}{\partial \alpha}\log L(\alpha, \hat{S}_1, \hat{S}_2)$$

$$= \int \frac{\partial}{\partial \alpha} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\}\, dH_n(c, \delta_1, \delta_2) = 0. \tag{4}$$

Empirically the proposed estimator $\hat{\alpha}$ can be obtained by iterating

$$\hat{\alpha}^{(k)} = \hat{\alpha}^{(k-1)} - \left\{\frac{\partial}{\partial \alpha} U(\alpha, \hat{S}_1, \hat{S}_2, H_n)\bigg|_{\alpha = \hat{\alpha}^{(k-1)}}\right\}^{-1} U(\hat{\alpha}^{(k-1)}, \hat{S}_1, \hat{S}_2, H_n),$$

where $\hat{\alpha}^{(k)}$ is the estimated value of α in the kth iteration.

2·3. *Asymptotic properties of the two-stage estimator*

When the nonparametric maximum likelihood estimators of S_j $(j = 1, 2)$ are used as the first-stage estimators for the marginals, it can be shown that $\hat{\alpha}$ is asymptotically normal under the required conditions. The result is formally stated in Theorem 1; the details of the proof are given in the Appendix. The main idea is that, by standard Taylor expansion techniques, one can write

$$0 = U(\hat{\alpha}, \hat{S}_1, \hat{S}_2, H_n)$$

$$= U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) + (\hat{\alpha} - \alpha_0)V(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) + O_p(|\hat{\alpha} - \alpha_0|^2),$$

where

$$V(\alpha_0, S_1, S_2, H) = \int \frac{\partial^2}{\partial \alpha^2} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\bigg|_{\alpha = \alpha_0} dH(c, \delta_1, \delta_2).$$

It will be shown that $V(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) \to V(\alpha_0, S_1, S_2, H) < 0$. If we express $U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n)$ as the sum of independently identically distributed components and smooth functionals of $\hat{S}_j - S_j$ $(j = 1, 2)$, asymptotic normality of $n^{\frac{1}{2}}U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n)$ can be established, and here asymptotic normality of $n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0)$ follows, where α_0 is the true value of α.

THEOREM 1. *Assume that the joint distribution of* (T_1, T_2) *follows a copula model in* (1) *with the true association parameter* $\alpha = \alpha_0$ *an interior point in the parameter space and that the following regularity conditions hold.*

(i) *The support of* S *is a bounded region* $[0, t_{01}] \times [0, t_{02}]$, G *has density* g *with respect to Lebesgue measure,* $G \ll F_1$ *and* $G \ll F_2$, *where* F_1 *and* F_2 *are the marginal distribution functions of* T_1 *and* T_2.

(ii) *We require that* $(\psi_1/g)\circ S_1^{-1}$ *and* $(\psi_2/g)\circ S_2^{-1}$ *are bounded and Lipschitz on* $[0, 1]$, *where* ψ_1 *and* ψ_2 *are derivatives of the influence curves of the marginals on the loglikelihood defined in* (A3), *where* $f \circ g$ *denotes the composite function of* f *and* g. *Specifically*

$$\psi_j(c) = \frac{\partial}{\partial \alpha} l_j\{\alpha, S_1(c), S_2(c)\}\bigg|_{\alpha = \alpha_0} h_{\alpha_0}(c) \quad (j = 1, 2), \tag{5}$$

$$l_1(\alpha, u_0, v_0) = \frac{\partial}{\partial u} l(\alpha, u, v_0)\bigg|_{u = u_0}, \quad l_2(\alpha, u_0, v_0) = \frac{\partial}{\partial v} l(\alpha, u_0, v)\bigg|_{v = v_0}. \tag{6}$$

(iii) *We require that*

$$I_1^{-1} = \int_0^{t_{01}} \frac{S_1(c)\{1-S_1(c)\}}{g(c)}\,\psi_1^2(c)\,dc < \infty,$$

$$I_2^{-1} = \int_0^{t_{02}} \frac{S_2(c)\{1-S_2(c)\}}{g(c)}\,\psi_2^2(c)\,dc < \infty.$$

(iv) *We require that*

$$\frac{\partial^3}{\partial\alpha^3} l\{\alpha, S_1(c), S_2(c)\}\Big|_{\alpha=\alpha_0}, \quad \frac{\partial^3}{\partial\alpha^2\,\partial u} l\{\alpha, u, S_2(c)\}\Big|_{\alpha=\alpha_0, u=S_1(c)},$$

$$\frac{\partial^3}{\partial\alpha^2\,\partial v} l\{\alpha, S_1(c), v\}\Big|_{\alpha=\alpha_0, v=S_2(c)}$$

are continuous and bounded for $(t_1, t_2) \in [0, t_{01}] \times [0, t_{02}]$.

Then $n^{\frac{1}{2}}(\hat{\alpha}-\alpha_0)$ *converges to a zero-mean normal random variable with variance equal to*

$$\sigma^2 = \operatorname{var}\{Q(\alpha_0, S_1, S_2, C, \delta_1, \delta_2)\},$$

where

$$Q(\alpha_0, S_1, S_2, c, \delta_1, \delta_2) = \frac{\partial}{\partial\alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\Big|_{\alpha=\alpha_0} + \sum_{j=1}^{2} [\delta_j - \{1-S_j(c)\}] \frac{\partial}{\partial\alpha} l_j\{\alpha, S_1(c), S_2(c)\}\Big|_{\alpha=\alpha_0} \frac{h_{\alpha_0}(c)}{g(c)}. \quad (7)$$

THEOREM 2. *Under the regularity conditions stated in Theorem* 1, *the variance* σ^2 *can be consistently estimated by the sample variance of*

$$Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, c, \delta_1, \delta_2) = \frac{\partial}{\partial\alpha} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\}\Big|_{\alpha=\hat{\alpha}} + \sum_{j=1}^{2} [\delta_j - \{1-\hat{S}_j(c)\}] \frac{\partial}{\partial\alpha} l_j\{\alpha, \hat{S}_1(c), \hat{S}_2(c)\}\Big|_{\alpha=\hat{\alpha}} \frac{h_{\hat{\alpha}}(c, \delta_1, \delta_2)}{g(c)};$$

that is

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \{Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, c_i, \delta_{1i}, \delta_{2i}) - \bar{Q}\}^2,$$

where

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^{n} Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, c_i, \delta_{1i}, \delta_{2i}).$$

Proof. Since $\hat{\alpha}$, $\hat{S}_1$ and $\hat{S}_2$ are uniformly consistent estimators for α, S_1 and S_2 respectively, and Q is continuous in α, S_1 and S_2,

$$\operatorname{var}\{Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, c, \delta_1, \delta_2)\} \to \operatorname{var}\{Q(\alpha_0, S_1, S_2, c, \delta_1, \delta_2)\}.$$

The sample variance of $Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, c, \delta_1, \delta_2)$ is then a consistent estimator of its variance. □

Remark 1. In our notation we omit the dependence of a quantity on the δ_j's in order to denote the summation of the same quantity over all possible (δ_1, δ_2) values. For example,

$$l\{\alpha_0, S_1(c), S_2(c)\}h_{\alpha_0}(c) = \sum_{\delta_1=0}^{1} \sum_{\delta_2=0}^{1} l\{\alpha_0, S_1(c), S_2(c), \delta_1, \delta_2\}h_{\alpha_0}(c, \delta_1, \delta_2).$$

Remark 2. Although the first condition in Theorem 1 requires that C be continuous and to have a density, we believe that the theory should still hold for the discrete case, and the assumption that $g(.)$ exists is just for convenience. To verify this argument, we ran a simulation, not presented here, with C generated from a discrete distribution and the result was still valid.

3. Numerical simulation

Simulations were carried out to examine the finite-sample performance of the proposed estimators of α and σ. Failure times (T_1, T_2) were generated from the Clayton family (Clayton, 1978), also known as the Gamma frailty model, with the survival function

$$S(s, t) = \{S_1(s)^{1-\alpha} + S_2(t)^{1-\alpha} - 1\}^{1/(1-\alpha)} \quad (\alpha > 1),$$

where $S_j(t) = \exp(-t)$ $(j = 1, 2)$ and $\tau = (\alpha - 1)/(\alpha + 1)$. The censoring variable C was generated from a uniform distribution. Define $\text{pr}(\delta = 1)$ as the prevalence level. The performance of the proposed estimators was evaluated under the combination of three dependence levels $(\tau = 0{\cdot}25, 0{\cdot}5, 0{\cdot}75)$, two prevalence levels (PL $\simeq 0{\cdot}2, 0{\cdot}5$) and two sample sizes $(n = 200, 400)$. We also computed an estimator $\tilde{\alpha}$ which solves $U(\alpha, S_1, S_2, H_n) = 0$. Note that $\tilde{\alpha}$ can be viewed as the 'best' estimator based on current status data since the marginals are completely specified. The difference between $\hat{\alpha}$ and $\tilde{\alpha}$ indicates the effect of using $\hat{S}_j$ instead of S_j $(j = 1, 2)$. The estimator of Kendall's tau is calculated using the relationship $q(\alpha) = \tau = (\alpha - 1)/(\alpha + 1)$. The results are summarised in Table 1, showing Monte Carlo estimates for the biases and standard deviations of $\hat{\alpha}$, $\tilde{\alpha}$, $\hat{\tau}$ and $\tilde{\tau}$ and the empirical coverage probability of the confidence interval, $(\hat{\alpha} - 1{\cdot}96\hat{\sigma}, \hat{\alpha} + 1{\cdot}96\hat{\sigma})$.

We can make several observations: the proposed procedure has reasonable performance when $n = 200$ and works better when $n = 400$; the estimating procedure usually works better under PL $\simeq 50\%$ than it does under PL $\simeq 20\%$; if we compare the two estimators $\tilde{\alpha}$ and $\hat{\alpha}$, the cost of using the marginal estimators seems to have more effect on the bias than on the standard deviation. We suggest that readers pay more attention to the performance of $\hat{\tau}$ than that of $\hat{\alpha}$ because τ has the same interpretation under all model alternatives. Note that, for the Clayton model, as $\alpha \to 1$, $\tau \to 0$, and, as $\alpha \to \infty$, $\tau \to 1$. If we use the delta method, $\sigma_{\hat{\tau}} \simeq q'(\alpha)\sigma = 2\sigma/(\alpha + 2)^2$, where $\sigma_{\hat{\tau}}$ is the standard deviation of $\hat{\tau}$. Therefore, although, as τ increases, the bias and variance of $\hat{\alpha}$ increase, the variance of $\hat{\tau}$ may still decrease.

The empirical coverage proabability of the proposed confidence interval is very close to the nominal level when $n = 400$. The bootstrap method can also be used to obtain a variance estimator. In a simulation with $n = 200$ we found that the empirical coverage probability using the bootstrap quantiles was close to the nominal level. Formal theoretical justification of the bootstrap method is beyond the scope of the paper, but it is covered by the general bootstrap theory in Politis & Romano (1994).

Note that the results in Table 1 exclude the very few cases where the estimates did not converge; when $n = 200$, $\tau = 0{\cdot}25$ and PL $\simeq 20\%$, there are 38 non-convergent cases out of 1000 and in the other scenarios fewer than 1% of cases failed to converge.

Table 1. *Simulation summary statistics of the two-stage estimator with* $n = 200$, 400 *for the Clayton model, based on* 1000 *simulations. The result for each estimator is the estimated bias, with estimated standard deviation in brackets*

Approx. PL	Estimator	$\tau = 0{\cdot}25$	$\tau = 0{\cdot}5$	$\tau = 0{\cdot}75$
			$n = 200$	
50%	$\tilde{\tau}$	0·001 (0·072)	0·002 (0·064)	0·002 (0·049)
	$\tilde{\alpha}$	0·024 (0·271)	0·009 (0·564)	0·421 (1·901)
	$\hat{\tau}$	0·010 (0·076)	0·022 (0·076)	0·032 (0·057)
	$\hat{\alpha}$	0·061 (0·293)	0·283 (0·698)	2·103 (5·700)
	Cover.	94·6%	95·3%	98·0%
20%	$\tilde{\tau}$	0·005 (0·110)	−0·005 (0·091)	−0·001 (0·049)
	$\tilde{\alpha}$	0·078 (0·428)	0·091 (0·763)	0·285 (1·757)
	$\hat{\tau}$	0·022 (0·113)	0·015 (0·098)	0·024 (0·055)
	$\hat{\alpha}$	0·150 (0·479)	0·304 (0·929)	1·514 (3·039)
	Cover.	97·1%	93·5%	98·0%
			$n = 400$	
50%	$\tilde{\tau}$	−0·001 (0·050)	0·000 (0·045)	0·000 (0·035)
	$\tilde{\alpha}$	0·005 (0·182)	0·030 (0·373)	0·149 (1·220)
	$\hat{\tau}$	−0·001 (0·053)	0·003 (0·050)	0·012 (0·040)
	$\hat{\alpha}$	0·008 (0·192)	0·063 (0·416)	0·656 (1·610)
	Cover.	95·5%	95·5%	95·2%
20%	$\tilde{\tau}$	−0·004 (0·085)	−0·001 (0·062)	0·000 (0·035)
	$\tilde{\alpha}$	0·018 (0·308)	0·056 (0·512)	0·165 (1·185)
	$\hat{\tau}$	0·005 (0·086)	0·008 (0·065)	0·014 (0·039)
	$\hat{\alpha}$	0·052 (0·322)	0·139 (0·563)	0·701 (1·480)
	Cover.	94·6%	94·8%	95·9%

PL, the prevalence level $\text{pr}(\delta = 1)$; cover., the empirical coverage probability of the confidence interval with endpoints $\hat{\alpha} \pm 1{\cdot}96\hat{\sigma}$.

4. AN ILLUSTRATIVE EXAMPLE

This 1991–93 study was designed to determine important risk factors for cardiovascular diseases in two towns, Chu-Dung and Pu-Tze, in Taiwan. The population of Taiwan consists of people of four major ancestral origins, namely Fukienese, Hakka, Chinese mainlander and aboriginal. Chu-Dung is a Hakka township located in the northern part of Taiwan and Pu-Tze is a Fukienese township located in the southwest of Taiwan. Five villages in each of the two townships were randomly selected from those with either more than 1000 people or population density greater than 200 per square kilometre. Altogether 6314 residents, including 2904 males and 3410 females, participated in the study. Of them 3824 were from Chu-Dung and 2490 from Pu-Tze.

Denote by (T_1, T_2, T_3) the ages of onset of hypertension, diabetes mellitus and hypercholesterolaemia respectively, and let C be the age at the monitoring time. Formally the prevalence of the diseases should have been determined based on doctors' diagnoses, but for convenience the three prevalence indicators were defined as follows: $\delta_1 = 1$ if the participant's systolic/diastolic blood pressures were at least 140/90 mmHg or he/she was taking medication for hypertension; $\delta_2 = 1$ if the participant's blood sugar was at least 126 mg/dl or he/she had a history of diabetes mellitus; and $\delta_3 = 1$ if the total cholesterol was at least 240 mg/dl. The data consist of $(C, \delta_1, \delta_2, \delta_3)$, where $\delta_j = I(T_j \leqslant C)$ $(j = 1, 2, 3)$.

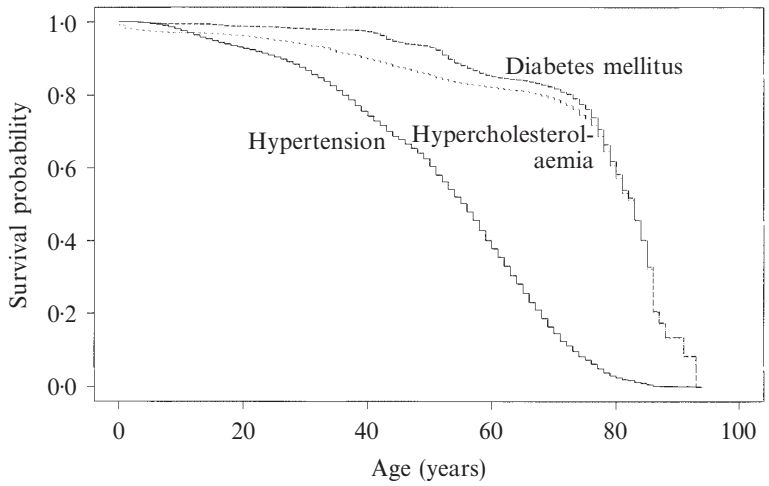

Fig. 1. Cardiovascular disease example. Nonparametric estimates of marginal survival functions for the three diseases.

The marginal survival functions, estimated by the nonparametric maximum likelihood estimator approach, are plotted in Fig. 1 which shows that the marginal behaviours of T_2 and T_3 are quite similar. The copula association parameters for the three bivariate models were estimated, under the assumption that (T_1, T_2), (T_1, T_3) and (T_2, T_3) all belong to the Clayton family; see Table 2. For each bivariate analysis, we also tested $H_0: \alpha = 1$ $(\tau = 0)$ versus $H_a: \alpha > 1$ $(\tau > 0)$ and the p-value of the test was computed. The ages of onset of hypertension and diabetes mellitus are significantly correlated, with $\hat{\tau}_{\mathrm{HT,DM}} = 0{\cdot}128$. Also the ages of onset of diabetes mellitus and that of hypercholesterolaemia are significantly correlated, with $\hat{\tau}_{\mathrm{DM,HC}} = 0{\cdot}304$. Note that $\hat{\tau}_{\mathrm{HT,HC}} = 0{\cdot}07$ with $p = 0{\cdot}052$, and that T_1 and T_3 are not correlated for females but are correlated for males. The disease relationships differ little between the two towns. Note that, although the analysis is based on the Clayton model assumption, model selection should not be an important issue here. According to Wang & Wells (2000), when τ is small all the copula models behave similarly. In fact when $\tau \to 0$ all the models approach the same one with $S(s, t) = S_1(s)S_2(t)$.

5. Discussion

Note that the dataset used in § 4 is not an ideal example since we implicitly made the dubious assumption that the one-time measurements could serve as formal medical diagnoses. Furthermore, the data were collected using a cluster-style sampling design, selecting participants by township. One therefore needs to be cautious in drawing conclusions about the general population based on the analysis; the results may be biased since the sample is not a simple random sample.

A nice feature of the proposed two-stage estimation procedure is that it can easily be adjusted under different scenarios. We have assumed univariate censorship, that is $C_1 = C_2 = C$. We believe that this setting is very common for current status data which often arise from survey studies where the two components are measured at the same time. However, our two-stage estimation procedure can still be extended to deal with bivariate current status data under different observation times C_1 and C_2. We can show that the

Table 2. *Summary statistics of the two-stage estimator applied to the dataset concerning three cardiovascular diseases in two towns under the Clayton* (1978) *model assumption. Tabulated are the estimated value of* τ, *the estimated value of* α, *the estimated value of* σ *and the p-value of the test of* $H_0: \alpha = 1$ $(\tau = 0)$ *versus* $H_\alpha: \alpha > 1$ $(\tau > 0)$

	HT versus DM	HT versus HC	DM versus HC
		Overall	
Estimate of τ	0·128	0·082	0·304
Estimate of α	1·295	1·178	1·875
Estimate of σ	0·155	0·110	0·175
p-value	0·028	0·052	0·000
		Females only	
Estimate of τ	0·228	0·055	0·266
Estimate of α	1·591	1·115	1·725
Estimate of σ	0·182	0·113	0·216
p-value	0·001	0·154	0·000
		Males only	
Estimate of τ	0·210	0·179	0·295
Estimate of α	1·531	1·436	1·836
Estimate of σ	0·179	0·160	0·265
p-value	0·002	0·003	0·001
		Chu-Dung only	
Estimate of τ	0·190	0·105	0·318
Estimate of α	1·469	1·234	1·932
Estimate of σ	0·147	0·110	0·221
p-value	0·001	0·017	0·000
		Pu-Tze only	
Estimate of τ	0·282	0·123	0·208
Estimate of α	1·785	1·280	1·527
Estimate of σ	0·242	0·162	0·254
p-value	0·001	0·042	0·003

HT, hypertension; DM, diabetes mellitus; HC, hypercholesterolaemia.

second stage estimator $\hat{\alpha}$ is still asymptotically normal at the rate $n^{\frac{1}{2}}$. However, variance estimation is much more difficult to examine analytically.

Additional covariate information may be incorporated to improve the first-stage estimation of S_j $(j = 1, 2)$ by accounting for heterogeneity of the sample; the observed data then consist of independently identically distributed replications of $\{C, \delta_1, \delta_2 Z\}$, where Z is the covariate. For a review of regression models for univariate current status data, see Huang & Wellner (1997). Under the specified regression model, $S_j(t)$ can be expressed as $S_j(t \mid \beta, z)$ and estimated by $\hat{S}_j(t \mid \hat{\beta}, z)$. Asymptotic normality of the resulting estimator of α at the usual $n^{\frac{1}{2}}$ rate can be established if one can show that with the new first-stage estimators $n^{\frac{1}{2}} b_{3n}$, where b_{3n} is defined in (A5), the estimator still converges to a zero-mean normal random variable. The asymptotic variance formula of $\hat{\alpha}$ will change accordingly. Similar arguments apply to the situation of dependent censoring, in which case the marginal estimators proposed by van der Laan & Robins (1998) may be used in the first-stage estimation.

In this paper, the marginals are estimated separately. If T_1 and T_2 are exchangeable, the marginals can be estimated jointly and the estimator $\hat{\alpha}$ may in turn be used to improve the marginal estimation. This issue deserves further investigation.

Bivariate current status data have the form of cross-sectional data which can be analysed using methods for two-by-two tables. For example, at an observed monitoring time c, define the odds ratio

$$\theta(c) = \frac{\mathrm{pr}(C = c, \delta_1 = \delta_2 = 1)\,\mathrm{pr}(C = c, \delta_1 = \delta_2 = 0)}{\mathrm{pr}(C = c, \delta_1 = 1, \delta_2 = 0)\,\mathrm{pr}(C = c, \delta_1 = 0, \delta_2 = 1)}.$$

Note that $\theta(c)$ measures the association between the prevalence indicators δ_1 and δ_2, and is sensitive to the marginal distributions of T_1 and T_2. However the proposed method evaluates the relationship between onset times of two diseases which are more related to the disease incidence.

Acknowledgement

We appreciate the valuable comments and suggestions of the three referees and the editor. Weijing Wang thanks Dr Wen-Harn Pan and Mr C. J. Yeh for providing the dataset and for their expert knowledge of cardiovascular epidemiology, Dr Chen-Hsin Chen for his comments and helpful suggestions and Ms Pei-Ching Wu for performing the data analysis. This work was partially funded by the National Science Council and Academia Sinica.

Appendix

Proof of Theorem 1

Stage (a): *Proof that* $V(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) \to V(\alpha_0, S_1, S_2, H) < 0$. From condition (iv), we can assume that the derivatives

$$\frac{\partial^3}{\partial\alpha^2\,\partial u} l(\alpha, u, v, \delta_1, \delta_2), \quad \frac{\partial^3}{\partial\alpha^2\,\partial v} l(\alpha, u, v, \delta_1, \delta_2)$$

are both bounded by some constant K on $[0, t_{01}] \times [0, t_{02}]$. We can write

$$\begin{aligned} V(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) &= V(\alpha_0, S_1, S_2, H_n) + \{V(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) - V(\alpha_0, S_1, S_2, H_n)\} \\ &= V(\alpha_0, S_1, S_2, H_n) + a_n. \end{aligned}$$

Since $\sup_{t \in [0, t_{0j}]} |\hat{S}_j(t) - S_j(t)| \to 0$ for $j = 1, 2$ (Groeneboom & Wellner, 1992, § 4.1), it follows that

$$\begin{aligned} |a_n| &\leqslant \int \left| \frac{\partial^2}{\partial\alpha^2} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\} - \frac{\partial^2}{\partial\alpha^2} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\} \right| dH_n(c, \delta_1, \delta_2) \\ &\leqslant \int \left| \frac{\partial^2}{\partial\alpha^2} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\} - \frac{\partial^2}{\partial\alpha^2} l\{\alpha, S_1(c), \hat{S}_2(c), \delta_1, \delta_2\} \right| dH_n(c, \delta_1, \delta_2) \\ &\quad + \int \left| \frac{\partial^2}{\partial\alpha^2} l\{\alpha, S_1(c), \hat{S}_2(c), \delta_1, \delta_2\} - \frac{\partial^2}{\partial\alpha^2} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\} \right| dH_n(c, \delta_1, \delta_2) \\ &\leqslant \int K |\hat{S}_1(c) - S_1(c)|\, dH_n(c, \delta_1, \delta_2) + \int K |\hat{S}_2(c) - S_2(c)|\, dH_n(c, \delta_1, \delta_2) \\ &\leqslant K \left(\sup_{0 \leqslant c \leqslant t_{01}} |\hat{S}_1(c) - S_1(c)| + \sup_{0 \leqslant c \leqslant t_{02}} |\hat{S}_2(c) - S_2(c)| \right) \\ &\to 0. \end{aligned}$$

Hence we have shown that $V(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) \to V(\alpha_0, S_1, S_2, H_n)$. By the Glivenko–Cantelli theorem, $H_n \to H$, and, by the Dominated Convergence Theorem, it follows that

$$V(\alpha_0, S_1, S_2, H_n) \to V(\alpha_0, S_1, S_2, H).$$

Finally, since $l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\} = \log\{h_\alpha(c, \delta_1, \delta_2)\}$,

$$\begin{aligned} V(\alpha_0, S_1, S_2, H) &= \frac{\partial^2}{\partial \alpha^2} \int dH_\alpha(c)\Big|_{\alpha=\alpha_0} - \int \left[\frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c)\}\right]^2\Big|_{\alpha=\alpha_0} dH_{\alpha_0}(c) \\ &= -\int \left[\frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c)\}\right]^2\Big|_{\alpha=\alpha_0} dH_{\alpha_0}(c) \\ &< 0. \end{aligned}$$

Stage (b): *Asymptotic distribution of* $n^{\frac{1}{2}} U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n)$. One can write

$$\begin{aligned} U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) = & \int \frac{\partial}{\partial \alpha} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\}\, dH_n(c, \delta_1, \delta_2) \\ = & \int \left\{\frac{\partial}{\partial \alpha} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\} - \frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\right\} d(H_n - H)(c, \delta_1, \delta_2) \\ & + \int \frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\, dH_n(c, \delta_1, \delta_2) \\ & + \int \left\{\frac{\partial}{\partial \alpha} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\} - \frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\right\} dH(c, \delta_1, \delta_2) \\ = & b_{1n} + b_{2n} + b_{3n}. \end{aligned} \tag{A1}$$

Since $\hat{S}_1 \to S_1$, $\hat{S}_2 \to S_2$, $n^{\frac{1}{2}}(H_n - H) = O_p(1)$ and

$$\frac{\partial}{\partial \alpha} l\{\alpha, \hat{S}_1(c), \hat{S}_2(c), \delta_1, \delta_2\}$$

is continuous and bounded, by the Dominated Convergence Theorem, the first term $b_{1n} = o_p(n^{-\frac{1}{2}})$.

The second term b_{2n} is a sum of independent and identically distributed quantities:

$$b_{2n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} l\{\alpha, S_1(c_i), S_2(c_i), \delta_{1i}, \delta_{2i}\}\Big|_{\alpha=\alpha_0}. \tag{A2}$$

Each term of (A2) has mean equal to

$$\int \frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\Big|_{\alpha=\alpha_0} dH(c, \delta_1, \delta_2) = 0$$

and variance equal to

$$\begin{aligned} & \int \left[\frac{\partial}{\partial \alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\Big|_{\alpha=\alpha_0}\right]^2 dH(c, \delta_1, \delta_2) \\ & \qquad = -\int \frac{\partial^2}{\partial \alpha^2} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\Big|_{\alpha=\alpha_0} dH(c, \delta_1, \delta_2). \end{aligned}$$

By the Central Limit Theorem, $n^{\frac{1}{2}} b_{2n}$ converges to a zero-mean normal random variable.

Applying von Mises expansions to b_{3n}, we obtain

$$b_{3n} = \int_0^{t_{01}} \text{IC}_1(t_1)\, d(\hat{S}_1 - S_1)(t_1) + \int_0^{t_{02}} \text{IC}_2(t_2)\, d(\hat{S}_2 - S_2)(t_2) + o_p(n^{-\frac{1}{2}}),$$

where $\text{IC}_j(t)$ is the influence curve of the functional $U(\alpha, S_1, S_2, H)$ at S_j $(j = 1, 2)$, obtained by differentiating $U\{\alpha, (1-\varepsilon_1)S_1 + \varepsilon_1\hat{S}_1, (1-\varepsilon_2)S_2 + \varepsilon_2\hat{S}_2, H\}$ with respect to ε_j and evaluating at $\varepsilon_1 = \varepsilon_2 = 0$:

$$\text{IC}_j(t) = -\int_0^{t_{0j}} \frac{\partial}{\partial\alpha} l_j\{\alpha, S_1(c), S_2(c)\}\Big|_{\alpha=\alpha_0} h_{\alpha_0}(c)\, dc. \tag{A3}$$

In the cases of no censoring or right censoring, $n^{\frac{1}{2}}(\hat{S}_j - S_j)$ can be written as sum of n independent and identically distributed terms, which would in turn imply the asymptotic normality of the term b_{3n}. For current status data, $(\hat{S}_j - S_j)$ has only rate of convergence $n^{1/3}$, and hence we cannot write them as independent and identically distributed sums directly. However, it is not possible to estimate smooth functionals of S_j at the rate $n^{1/2}$. We shall show that b_{3n} is a sum of such smooth functionals satisfying the conditions in Huang & Wellner (1995) and hence is asymptotically normally distributed.

To do this define the functionals

$$v_j(S) = \int_0^{t_{0j}} \text{IC}_j(x)\, dS(x) = -\int_0^{t_{0j}} \Psi_j(x)\, dS(x) \quad (j = 1, 2),$$

where

$$\Psi_j(x) = -\text{IC}_j(x) = \int_0^x \psi_j(c)\, dc.$$

Then $b_{3n} = v_1(\hat{S}_1) - v_1(S_1) + v_2(\hat{S}_2) - v_2(S_2)$. If conditions (i), (ii) and (iii) in Theorem 1 hold, according to Huang & Wellner (1995),

$$\begin{aligned} n^{\frac{1}{2}} \int_0^{t_{0j}} \text{IC}_j(t_j)\, d(\hat{S}_j - S_j)(t_j) &= n^{\frac{1}{2}}\{v_j(\hat{S}_j) - v_j(S_j)\} = -n^{\frac{1}{2}}(P_n - P)(\tilde{l}) + o_p(1) \\ &= -n^{-\frac{1}{2}} \sum_{i=1}^n \{\tilde{l}(c_i, \delta_{ji}, S_j, G, \psi_j) - E_{S_j,G}(\tilde{l})\} + o_p(1), \end{aligned}$$

where

$$\tilde{l}(c, \delta, S, G, \psi) = -[\delta - \{1 - S(c)\}]\frac{\psi(c)}{g(c)} I\{g(c) > 0\}. \tag{A4}$$

Therefore

$$b_{3n} = -\frac{1}{n}\sum_{i=1}^n \{\tilde{l}(c_i, \delta_{1i}, S_1, G, \psi_1) + \tilde{l}(c_i, \delta_{2i}, S_2, G, \psi_2) - E_{S_1,G}(\tilde{l}) - E_{S_2,G}(\tilde{l})\} + o_p(n^{-\frac{1}{2}}). \tag{A5}$$

Combining (A2) and (A5) in (A1), we obtain that

$$U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n) = \frac{1}{n}\sum_{i=1}^n Q(\alpha_0, S_1, S_2, c_i, \delta_{1i}, \delta_{2i}) - E(Q) + o_p(n^{-\frac{1}{2}}),$$

where

$$Q(\alpha_0, S_1, S_2, c, \delta_1, \delta_2) = \frac{\partial}{\partial\alpha} l\{\alpha, S_1(c), S_2(c), \delta_1, \delta_2\}\Big|_{\alpha=\alpha_0} - \tilde{l}(c, \delta_1, S_1, G, \psi_1) - \tilde{l}(c, \delta_2, S_2, G, \psi_2),$$

and $E(Q)$ denotes the expected value of Q.

Note that (5) and (A4) imply that

$$\tilde{l}(c_i, \delta_{ji}, S_j, G, \psi_j) = -[\delta_{ji} - \{1 - S_j(c_i)\}] \frac{\partial}{\partial \alpha} l_j\{\alpha, S_1(c_i), S_2(c_i)\}\Big|_{\alpha=\alpha_0} \frac{h_{\alpha_0}(c)}{g(c)},$$

where

$$\frac{h_{\alpha_0}(c, \delta_1, \delta_2)}{g(c)} = \delta_1\delta_2 S_{11}(c) + (1-\delta_1)\delta_2 S_{01}(c) + \delta_1(1-\delta_2)S_{10}(c) + (1-\delta_1)(1-\delta_2)S_{00}(c)$$

is independent of g. Hence Q does not depend on G.

Therefore, by the Central Limit Theorem, $n^{\frac{1}{2}}U(\alpha_0, \hat{S}_1, \hat{S}_2, H_n)$ is asymptotically normal with mean zero and variance equal to

$$\sigma^2 = \text{var}\{Q(\alpha_0, S_1, S_2, c, \delta_1, \delta_2)\},$$

where Q is given in (7).

REFERENCES

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. & SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete observations. *Ann. Math. Statist.* **26**, 641–7.

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**, 141–51.

DIAMOND, I. D. & MCDONALD, J. W. (1992). The analysis of current status data. In *Demographic Applications of Event History Analysis*, Ed. J. Trussell, R. Hankinson and J. Tilton, pp. 231–52. New York: Oxford University Press.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845–54.

GENEST, C., GHOUDI, K. & RIVEST, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543–52.

GENEST, C. & MACKAY, R. J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *Am. Statistician* **40**, 280–3.

GENEST, C. & RIVEST, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Am. Statist. Assoc.* **88**, 1034–43.

GROENEBOOM, P. & WELLNER, J. A. (1992). *Information Bounds and Non-Parametric Maximum Likelihood Estimation.* Boston: Birkhäuser.

HSU, L. & PRENTICE, R. L. (1996). On assessing the strength of dependency between failure time variates. *Biometrika* **83**, 491–506.

HUANG, J. & WELLNER, J. A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, Case 1. *Statist. Neer.* **49**, 153–63.

HUANG, J. & WELLNER, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Ed. D. Y. Lin and T. R. Fleming, pp. 123–69. New York: Springer-Verlag.

JEWELL, N. P., MALANI, H. M. & VITTINGHOFF, E. (1994). Nonparametric estimator for a form of doubly censored data with application to two problems in AIDS. *J. Am. Statist. Assoc.* **89**, 7–18.

JEWELL, N. P. & SHIBOSKI, S. C. (1990). Statistical analysis of HIV infectivity based on partner study data. *Biometrics* **46**, 1133–50.

JEWELL, N. P. & VAN DER LAAN, M. (1997). Singly and doubly censored current status data with extensions to multistate counting processes. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Ed. D. Y. Lin and T. R. Fleming, pp. 171–84. New York: Springer-Verlag.

KLEIN, R. W. & SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–421.

LIN, D. Y., OAKES, D. & YING, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289–98.

OAKES, D. (1989). Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**, 487–93.

PETO, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* **22**, 86–91.

POLITIS, D. N. & ROMANO, J. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031–50.

RABINOWITZ, D., TSIATIS, A. & ARAGON, J. (1995). Regression with interval censored data. *Biometrika* **82**, 501–13.

ROSSINI, A. J. & TSIATIS, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Am. Statist. Assoc.* **91**, 713–21.

SHIH, J. H. & LOUIS, T. A. (1995). Inference on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384–99.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data. *J. R. Statist. Soc.* B **38**, 290–5.

VAN DER LAAN, M. J. & ROBINS, J. M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *J. Am. Statist. Assoc.* **93**, 693–701.

WANG, W. & WELLS, M. T. (1997). Nonparametric estimators of the bivariate survival function under simplified structures. *Biometrika* **84**, 863–80.

WANG, W. & WELLS, M. T. (2000). Model selection and semiparametric inference for bivariate failure time data. *J. Am. Statist. Assoc.* **95**, 62–72.

[*Received November* 1998. *Revised June* 2000]