Communications for Statistical Applications and Methods
2015, Vol. 22, No. 6, 599–604

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.6.599
Print ISSN 2287-7843 / Online ISSN 2383-4757

Estimation of Conditional Kendall's Tau for Bivariate Interval Censored Data

Yang-Jin Kim[1,a]

[a]Department of Statistics, Sookmyung Women's University, Korea

Abstract

Kendall's tau statistic has been applied to test an association of bivariate random variables. However, incomplete bivariate data with a truncation and a censoring results in incomparable or unorderable pairs. With such a partial information, Tsai (1990) suggested a conditional tau statistic and a test procedure for a quasi independence that was extended to more diverse cases such as double truncation and a semi-competing risk data. In this paper, we also employed a conditional tau statistic to estimate an association of bivariate interval censored data. The suggested method shows a better result in simulation studies than Betensky and Finkelstein's multiple imputation method except a case in cases with strong associations. The association of incubation time and infection time from an AIDS cohort study is estimated as a real data example.

Keywords: AIDS, bivariate interval censored data, conditional Kendall's tau, jackknife variance, quasi-independence, unorderable pairs

1. Introduction

There are several approaches to measure associations between two variables. Among them, Kendall's tau is commonly used because of rank invariant property and has powerful asymptotic properties such as a U-statistic. With a pair of bivariate data (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}), a Kendall's tau is defined as $\tau = E\{\text{sgn}((T_{1i} - T_{1j})(T_{2i} - T_{2j}))\}$ where $\text{sgn}(x) = 1$ for $x > 0$, $\text{sgn}(x) = -1$ for $x < 0$, and $\text{sgn}(x) = 0$ for $x = 0$. It is also interpreted as the difference between concordance rate and discordance rate. The value of τ is between -1 and 1. When T_1 and T_2 are independent, the probabilities of concordance and discordance are same and implies $\tau = 0$. For complete data, it is estimated as,

$$\hat{\tau} = \frac{\sum_{1 \le i \le j \le n} a_{ij} b_{ij}}{\binom{n}{2}}, \tag{1.1}$$

where a score $a_{ij} = 1$ if $T_{1i} > T_{1j}$; -1 if $T_{1i} < T_{1j}$ and $b_{ij} = 1$ if $T_{2i} > T_{2j}$; -1 if $T_{2i} < T_{2j}$. Then $a_{ij}b_{ij} = 1$ for $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0$ which is a concordant pair and $a_{ij}b_{ij} = -1$ for $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0$ which is a discordant pair. Then $\hat{\tau}$ is an unbiased estimator of τ and $n^{1/2}(\hat{\tau} - \tau)$ is known to have asymptotically normal distribution (Hoeffding, 1948).

However, for an incomplete data with censoring and truncation, the estimation of τ has been challenging. For a right censored data, in order to compensate incomparable pairs, Brown *et al.* (1974)

This research was supported by the Sookmyung Women's University research grants 2015.

[1] Department of Statistics, Sookmyung Women's University, Chengpa-Dong, Yonnsan-Gu, Seoul 140-742, Korea.
E-mail: yjin@sookmyung.ac.kr

Published 30 November 2015 / journal homepage: http://csam.or.kr

suggested weighted scores. Weier and Basu (1980) modified the estimator to reflect the admissible information of comparable pairs. Oakes (1982) suggested an estimator using only comparable pair and tested a null hypothesis, $H_0 : T_1 \perp T_2$. Wang and Wells (2000) utilized an integral form with an estimated bivariate survival function and expressed it as a V-statistic. Hsieh (2010) considered several types of imputed times to replace a censored data and calculated τ estimates based on imputed values.

Tsai (1990) has alternatively extended Oakes' idea and considered a comparable pair for a left truncation data. In more detail, when T_1 and T_2 are denoted as left truncation time and failure time, respectively, the comparable pair satisfies $T_1 < T_2$. For this restricted data, he suggested a conditional tau statistic in that uses only partial information. A comparable pair with a left truncation is then defined as

$$\Omega_{ij} = \left\{\max\left(T_{1i}, T_{1j}\right) \leq \min\left(T_{2i}, T_{2j}\right)\right\} \tag{1.2}$$

and a corresponding conditional tau is estimated as

$$\hat{\tau}_c = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{sgn}\left(\left(T_{1i} - T_{1j}\right)\left(T_{2i} - T_{2j}\right)\right) I\left(\Omega_{ij}\right)}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I\left(\Omega_{ij}\right)}.$$

This restriction also derives a test statistic for a quasi-independence, $H_0 : T_1 \perp_Q T_2$. Martin and Betensky (2005) consider more general truncation schemes and suggest several estimators for a conditional tau. They represented these estimators with U-statistics and derived corresponding variances. For example, the available data includes censoring information when a right censoring data is added to a left truncation, C_i, C_j, $\delta_i = I(T_{2i} < C_i)$ and $\delta_j = I(T_{2j} < C_j)$. Then a comparable or orderable pair (1.2) is redefined with $Y_i = \min(T_{2i}, C_i)$ and $Y_j = \min(T_{2j}, C_j)$,

$$A_{ij} = \left\{\max\left(T_{1i}, T_{1j}\right) \leq \min\left(Y_i, Y_j\right)\right\} \cap \left\{\delta_i\delta_j = 1 \vee \delta_i \operatorname{sgn}\left(Y_j - Y_i\right) = 1 \vee \delta_j \operatorname{sgn}\left(Y_i - Y_j\right) = 1\right\} \tag{1.3}$$

and a corresponding conditional tau is estimated by

$$\tilde{\tau}_c = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{sgn}\left(\left(Y_i - Y_j\right)\left(T_{1i} - T_{1j}\right)\right) I\left(A_{ij}\right)}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I\left(A_{ij}\right)}.$$

They also extended to test a quasi-independence for doubly truncation data and an interval censored data with a truncation. However, the suggested estimators can make poor performance when a true τ gets away from zero. To correct a selection bias from censoring, Lakhal *et al.* (2009) suggested an IPCW estimator under a bivariate right censoring without truncation. Under a semi-competing risk data, Hsieh and Huang (2015) applied the IPCW technique to estimate a conditional tau. Denote X_i and Y_i as time to non-terminal event and time to terminal event, respectively. Therefore, X_i is censored by Y_i but does not censor Y_i. Define the observable variables $Z_i = \min(X_i, Y_i, C_i)$, $T_i = \min(Y_i, C_i)$, $\delta_{X_i} = I(X_i \leq Y_i \wedge C_i)$ and $\delta_{Y_i} = I(Y_i \leq C_i)$. Then a comparable and orderable pair is defined as

$$S_{ij} = \left\{\min\left(X_i, X_j\right) \leq \min\left(Y_i, Y_i\right)\right\} \cap \left\{\delta_{X_i}\delta_{X_j} = 1 \vee \delta_{X_i} \operatorname{sgn}\left(Y_j - Y_i\right) = 1 \vee \delta_{X_j} \operatorname{sgn}\left(Y_i - Y_j\right) = 1\right\}, \tag{1.4}$$

and a corresponding conditional tau is

$$\hat{\tau}_{c,w} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{sgn}\left(\left(Y_i - Y_j\right)\left(Z_i - Z_j\right)\right) I\left(S_{ij}\right)/\hat{p}_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I\left(S_{ij}\right)/\hat{p}_{ij}},$$

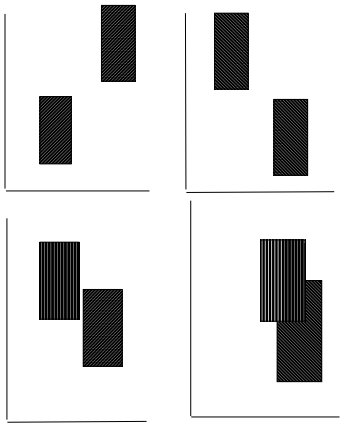

Figure 1: *Bivariate interval censored data.*

where and $\hat{p}_{ij} = [\hat{P}(\min(C_i, C_j) \geq \min(X_i, X_j))]^2$. To estimate $\hat{P}(C > t)$, the Kaplan-Meier estimator is implemented with $(Y_i, 1 - \delta_i)$. Our interest in this paper is to estimate a conditional tau under bivariate interval censored data and test a quasi-independence between two interval censored data. In Section 2, we extend a conditional tau to bivariate interval censored data. In Section 3, finite sample properties of the suggested method are evaluated through some simulation studies and AIDS cohort study is analyzed in Section 4. We remark some issues in Section 5.

2. Conditional Kendall's Taus for Interval Censored Data

We propose a new estimator for τ_c under bivariate interval censored data where it is more burdensome to calculate both concordance and discordance because of overlapping rectangles and of uncertain orderings of even non-overlapping rectangles. Figure 1 shows a graphic representing bivariate interval censored data with concordant, discordant and incomparable pairs. Top left is a concordant pair, top right is a discordant, bottom left is non-overlapping but incomparable and bottom right shows an overlapping case. There are several approaches to calculate an association for interval censored data. Betensky and Finkelstein (1999) implemented a multiple imputation technique using the estimated bivariate distribution in order to replace unknown bivariate failure times. However, the derived imputed results reproduce rectangles which can result in incomparable pairs. Therefore, these are again excluded from the calculation. As another method, Bogaerts and Lesaffre (2008) utilized the following formulae, $\tau = 4 \iint F(x, y) dF(x, y) - 1$, where a bivariate distribution function was estimated with a smoothing technique based on a normal approximation. They also suggested a cross-hazard function to measure a local association; however, diverse simulation studies were not discussed in their paper.

Denote $\{(L_i^1, R_i^1), (L_i^2, R_i^2), \delta_{1i}, \delta_{2i}, i = 1, \ldots, n\}$ as an observable data consisting of bivariate interval censored data and censoring indicators, respectively. Here, instead of observing failure times T_{1i} and T_{2i}, we observe interval censoring times satisfying $L_i^1 < T_{1i} < R_i^1$ and $L_i^2 < T_{2i} < R_i^2$, respectively. Censoring indicators are also defined as $\delta_{1i} = (L_i^1 < R_i^1 < \infty)$, that is, $\delta_{1i} = 0$ if $R_{i1} = \infty$. Similar definition is defined with δ_{2i}. We assume that interval censoring times are independent of both failure times. In order to extend the comparable and orderable pair like (1.2), (1.3) and (1.4) to a bivariate interval censored data, define $\tilde{L}_{ij}^1 = \max(L_i^1, L_j^1)$, $\tilde{L}_{ij}^2 = \max(L_i^2, L_j^2)$, $\tilde{R}_{ij}^1 = \min(R_i^1, R_j^1)$ and $\tilde{R}_{ij}^2 = \min(R_i^2, R_j^2)$. Furthermore, pairwise censoring indicators are defined as $\delta_{ij}^1 = 1 - (1 - \delta_{1i})(1 - \delta_{1j})$ and

$\delta_{ij}^2 = 1 - (1 - \delta_{2i})(1 - \delta_{2j})$. Then the comparable and orderable set is

$$B_{ij} = \left\{\left(\tilde{R}_{ij}^1 < \tilde{L}_{ij}^1\right) \cap \left(\tilde{R}_{ij}^2 < \tilde{L}_{ij}^2\right)\right\}$$

a conditional tau is calculated as

$$\hat{\tau}_c^B = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{sgn}\left(\left(\tilde{R}_{ij}^1 - \tilde{L}_{ij}^1\right)\left(\tilde{R}_{ij}^2 - \tilde{L}_{ij}^2\right)\right) I\left(B_{ij}\right)}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I\left(B_{ij}\right)}.$$

The jackknife approach is applied since the asymptotic variance of $\hat{\tau}_c^B$ is difficult to derive,

$$\hat{\sigma}_c^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\tau}_c^{B(i)} - \hat{\tau}_c^{B(\cdot)}\right)^2,$$

where $\hat{\tau}_c^{B(i)}$ is an estimate based on the data deleting the i^{th} subject and $\hat{\tau}_c^{B(\cdot)} = n^{-1}\sum_{i=1}^{n}\hat{\tau}^{B(i)}$.

3. Simulation

In this section, simulation studies are performed to examine the finite sample performance of the suggested estimators. 300 simulation samples were generated and bivariate failure times are generated from Clayton's copula model where the association parameter can be expressed as $\alpha = 2\tau/(1-\tau)$, that is, (T_1, T_2) has a following joint survival function,

$$S(t_1, t_2) = \left(S_1(t_1)^{-\alpha} + S_2(t_2)^{-\alpha} - 1\right)^{-\frac{1}{\alpha}}.$$

In order to make an interval censored data for T_1, generate m's u from a uniform distribution $U(0, 0.15)$. Then set $L_i^1 = \sum_{k=1}^{j} u_j = \tilde{u}_j$ and $R_i^1 = L_i^1 + u = \tilde{u}_{j+1}$ satisfying $L_i^1 < T_1 < R_i^1$. Here m is set depending on right censoring rate. Similar method is applied to make an interval censored data for T_2. With each setup described above, two sample sizes ($n = 100, 200$) are considered. In this simulation, we set $\tau = 0.0, 0.3, 0.5, 0.7$. Where, $\tau = 0$ means T_1 and T_2 are independent. We consider two different high right censoring rates since it is related with the amount of orderable pairs. Table 1 shows the simulation result under moderate right censoring rate and Table 2 is one for a high right censoring rate. Under a moderate right censoring rate, the suggested method brings in unbiased estimators and satisfies 95% coverage probabilities (CP) at most cases except $\tau = 0.7$. The averaged standard error based on the Jackknife method (**SSE = $\hat{\sigma}_c$**) is also close to the empirical standard deviation (SEE). However, a high right censoring case (i.e. a high rate of unorderable pair) provides under-estimated results except $\tau = 0$. In particular, a stronger association produces a more underestimated estimation. Table 3 shows biases and MSEs under two different censoring rates that compares the suggested method and the Betensky and Finkelstein's estimator. Under moderate censoring, the suggested estimate has a smaller bias and smaller MSE but at a high right censoring, the conditional tau has larger bias at $\tau = 0.5$.

4. Data Analysis

The suggested method is applied to the datasets from AIDS Clinical Trials Group Protocol 181. A cohort study of 257 individuals with Type A or Type B hemophilia was analyzed by DeGruttola and Lagakos (1989) and Kim *et al.* (1993) to estimate the distribution and regression coefficient,

Table 1: Estimation of τ_c^B with moderate right censored data

τ	$n = 100$				$n = 200$			
	$\hat{\tau}_c^B$	SSE($\hat{\tau}_c^B$)	SEE($\hat{\tau}_c^B$)	CP	$\hat{\tau}_c^B$	SSE($\hat{\tau}_c^B$)	SEE($\hat{\tau}_c^B$)	CP
0.0	−0.006	0.078	0.079	0.950	−0.006	0.057	0.055	0.930
0.3	0.293	0.074	0.075	0.937	0.296	0.051	0.052	0.940
0.5	0.503	0.061	0.061	0.943	0.498	0.041	0.043	0.960
0.7	0.717	0.033	0.035	0.957	0.689	0.031	0.032	0.913

Table 2: Estimation of τ_c^B with high right censored data

τ	$n = 100$				$n = 200$			
	$\hat{\tau}_c^B$	SSE($\hat{\tau}_c^B$)	SEE($\hat{\tau}_c^B$)	CP	$\hat{\tau}_c^B$	SSE($\hat{\tau}_c^B$)	SEE($\hat{\tau}_c^B$)	CP
0.0	0.003	0.078	0.080	0.947	0.006	0.056	0.056	0.956
0.3	0.277	0.074	0.075	0.957	0.279	0.051	0.051	0.933
0.5	0.466	0.063	0.062	0.933	0.471	0.040	0.042	0.923
0.7	0.666	0.036	0.038	0.913	0.662	0.027	0.028	0.750

Table 3: Bias and MSE of τ_c^B and τ_m at moderate right and high right censored case

		$\tau = 0.0$		$\tau = 0.3$		$\tau = 0.5$	
		Bias	MSE	Bias	MSE	Bias	MSE
Moderate right censored	$\hat{\tau}_c^B$	0.006	0.009	0.007	0.013	0.003	0.009
	$\hat{\tau}_m$	0.065	0.023	0.077	0.021	0.013	0.010
High right censored	$\hat{\tau}_c^B$	0.003	0.005	0.023	0.006	0.034	0.005
	$\hat{\tau}_m$	0.060	0.013	0.039	0.007	0.009	0.006

respectively. As the application of the suggested method, we investigate the association between HIV infection time and AIDS incubation time (or diagnosis time) using a null hypothesis, $H_0 : \tau_c = 0$.

HIV infection is determined from a blood test; therefore, the exact HIV infection time is unavailable and is observed with an interval form such as $XL \le X \le XR$. Here, XL is the last inspection time with negative infection and XR is the first inspection time with positive infection, respectively. Therefore, a corresponding incubation time is also an interval censored with (TL, TR) where T denotes an AIDS diagnosis time, then $TL = T - XR$ and $TR = T - XL$, respectively. Among heavily treated 97 HIV positive patients, 29 patients got AIDS positive results. When applying the suggested method, a conditional tau $\tau_c^B = -0.189$ (se $= \hat{\sigma}_c = 0.064$) is estimated and a corresponding p-value $= 0.012$ is derived based on an asymptotic normality. It indicates that infection time and incubation time are negatively correlated and patients with an early infection time have a higher chance to extend the AIDS diagnosis time than patients with a late infection time.

5. Discussion

In this paper, we investigate an association measure for bivariate interval censored data. However, incomparable pairs occur and an ordinary tau statistic is not suitable due to the existence of both overlapping rectangles and uncertain ordering for even non-overlapping rectangles. Tsai (1990) suggested a quasi-independence based on the comparable set for truncated data and Martin and Betensky (2005) applied it to data with doubly truncation and interval censored with truncation. We defined an orderable set and calculated a conditional tau by extending their method to bivariate interval censored data.

According to simulation results, the suggested method provides unbiased result under a moderate right censoring and a jackknife variance estimator works well for several tau values. However, extra simulation study shows the conditional tau estimator results in under-estimation under a high right

censoring. In order to solve this problem, we can consider a weighting method following Hsieh and Huang (2015) who considered a semi-competing risk data. However, while their weight is based on the distribution of right censoring times, our case is not so simple since interval censoring data occurs with observation process; consequently, modeling such process requires several assumptions and models. Finkelstein *et al.* (2002) derived a joint likelihood of observation process and event process and applied an EM algorithm in order to estimate related parameters. Therefore, the estimated observation process can be implemented to reflect a high right censoring rate. Another alternative weighting method is related with the recovery of the loss of information. Bivariate interval censoring data has pairs with overlapping rectangles and uncertain orderings of even non-overlapping rectangles. Therefore, it would bring a more efficient result by assigning some weight instead of dropping these pairs from the estimation procedure. We also consider to extend Brown *et al.* (1974)'s method.

References

Betensky, R. and Finkelstein, D. F. (1999). An extension of Kendall's coefficient of concordance to bivariate interval censored data, *Statistics in Medicine*, **18**, 3101–3109.

Bogaerts, K. and Lesaffre, E. (2008). Estimating local and global measures of association for bivariate interval censored data with a smooth estimate of density, *Statistics in Medicine*, **27**, 5941–5955.

Brown, B. W., Hollander, M. and Korwar, R. M. (1974). *Nonparametric Tests of Independence for Censored Data, with Applications to Heart Transplant Studies*, Reliability and Biometry, 327–354.

DeGruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, **45**, 1–12.

Finkelstein, D. M., Giggins, W. B. and Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring, *Biometrics*, **58**, 298–304.

Hoeffding, W. (1948). A class of statistics with asymptotic normal distributions, *The Annals of Mathematical Statistics*, **19**, 293–325.

Hsieh, J-J. (2010). Estimation of Kendall's tau from censored data, *Computational Statistics and Data Analysis*, **54**, 1613–1621.

Hsieh, J. J. and Huang, W. C. (2015). Nonparametric estimation and test of conditional Kendall's tau under semi-competing risk data and truncated data, *Journal of Applied Statistics*, **42**, 1602–1616.

Kim, M. Y., De Gruttola, V. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS, *Biometrics*, **49**, 13–22.

Lakhal, L., Rivest, L. P. and Beaudoin, D. (2009). IPCW estimator for Kendall's tau under bivariate censoring, *International Journal of Biostatistics*, **5**, 1–20.

Martin, E. C. and Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times bias conditional Kendall's tau, *Journal of American Statistical Association*, **100**, 484–492.

Oakes, D. (1982). A concordance test for independence in the presence of censoring, *Biometrics*, **38**, 451–455.

Tsai, T. W (1990). Testing the assumption of independence of truncation time and failure time, *Biometrika*, **77**, 169–177.

Wang, W. and Wells, M. (2000). Estimation of Kendall's tau under censoring, *Statistica Sinica*, **10**, 1199–1215.

Weier, D. R. and Basu, A. P. (1980). An investigation of Kendall's τ modified for censored data with applications, *Journal of Statistics and Planning Inference*, **4**, 381–390.

Received July 13, 2015; Revised October 2, 2015; Accepted October 6, 2015