The International Journal of Biostatistics

Volume 3, Issue 1

2007

Article 8

The Analysis of Bivariate Truncated Data Using the Clayton Copula Model

Antai Wang, *Georgetown University*

Recommended Citation:

Wang, Antai (2007) "The Analysis of Bivariate Truncated Data Using the Clayton Copula Model," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 8.

DOI: 10.2202/1557-4679.1050

The Analysis of Bivariate Truncated Data Using the Clayton Copula Model

Antai Wang

Abstract

In individuals infected with human immunodeficiency virus (HIV), distributions of quantitative HIV RNA measurements may be highly left-censored due to values falling below assay detection limits (DL). It is of the interest to find the relationship between plasma and semen viral loads. To address this type of problem, we developed an empirical goodness-of-fit test to check the Clayton model assumption for bivariate truncated data. We also used truncated tau to estimate the dependence parameter in the Clayton model for this type of data. It turns out that the proposed methodology works for both truncated and fixed left censored bivariate data. The proposed test procedure is demonstrated using an HIV data set, and statistical inference is drawn based on corresponding test result.

KEYWORDS: copula models, goodness-of-fit tests, the clayton model, bivariate truncated data

Author Notes: The author would like to express gratitude to three referees for a thorough review and constructive comments that led to some substantial improvements in this paper. The author is especially thankful to Dr. Lei Nie for valuable discussions and for sending him the HIV data set, as well as Dr. Edmund A. Gehan for helpful comments regarding this paper.

1 Introduction

In individuals infected with human immunodeficiency virus (HIV), distributions of quantitative HIV RNA measurements may be highly left-censored due to values falling below assay detection limits (DL). It is of scientific interest to find the relationship between plasma and semen viral loads. Bivariate normal distributions may be utilized to model the association. However, due to the large proportion of data falling below DL, it is not straightforward to verify the normality assumption, sometimes, the normality assumption may not even hold. The purpose of this paper is to propose an alternative way to analyze bivariate truncated data using the Clayton copula model. Although the proposed strategy can also be applied to bivariate data subject to fixed left censoring (see section 4), to make our presentation clearer, we only consider the truncated data situation in the following methodology development.

To simplify our data analysis, we assume that the original (untruncated or uncensored) data follow some absolutely continuous copula model. According to Sklar's Theorem (1959), the joint survivor function of a bivariate random vector (T_1, T_2) with continuous marginal survivor functions $S_k(u_k) = \text{pr}(T_k > t_k)$ for $k = 1, 2$ can be uniquely determined by its associated copula or dependence function, C , defined for all $(t_1, t_2) \in [0, 1]^2$ by

$$S(t_1, t_2) = C(S_1(t_1), S_2(t_2)).$$

As pointed out by Genest, Ghouli and Rivest (1995) many multivariate dependence models can be generated by parametric families (C_θ) of copulas, where θ is the dependence parameter of the copula model. Copula models have been widely used in that their dependence structure can be separated from the marginal distributions.

Besides the normal copula family, one of the most popular families of the copula models is the Archimedean copula family. Archimedean copula C_θ is defined as:

$$C_\theta(t_1, t_2) = p\{q(t_1) + q(t_2)\}$$

for some convex, decreasing function q on $(0, 1]$ with $q(1) = 0$ and $p(\cdot)$ is specified as the inverse function of q . Archimedean copula models have wide applications in multivariate survival analysis. Some well-known models such as the Clayton model (Clayton 1978), the Hougaard model (Hougaard 1986) and the Frank model (Genest 1987) all belong to this family.

In this paper, we have considered the application of the Clayton model to the analysis of bivariate data under left truncation or fixed left censoring. The

proposed methodologies are based on the fact that the copula associated with the Clayton model is the only absolutely continuous copula that is preserved under bivariate truncation (Oakes 2005). From this fact, we can deduce that if the Clayton model assumption is valid, the truncated sample must follow the Clayton model. Conversely, assuming the absolutely continuous copula model for original data and the Clayton model for truncated data, we can infer that the Clayton model assumption still holds for original data. A goodness-of-fit test of the Clayton model assumption for original data can thus be developed based on truncated sample, i.e., among pairs (T_1, T_2) with both T_1 and T_2 being observed. Furthermore, if the proposed test justifies the Clayton model assumption, one can evaluate the association parameter θ for original data using truncated τ . These are the objectives of this paper: in section 2, we have described a goodness-of-fit test procedure for original bivariate data based on truncated data. In section 3, we propose the use of truncated τ as a parameter estimator and evaluate its performance by simulations. In section 4, we demonstrate the proposed methodologies using a fixed left censored HIV data set. The paper ends with some discussions in section 5.

2 A Goodness-of-Fit Test

Let (T_{1i}, T_{2i}) ($i = 1, \dots, n$) denote the dependent pairs whose joint distribution is generated by some absolutely continuous copula C_θ and (L_1, L_2) the constant truncation vector. Also assume that (T_{1i}, T_{2i}) are random samples with continuous marginal survivor functions so that the joint survivor function $S(t_1, t_2) = C_\theta\{S_1(t_1), S_2(t_2)\}$ is also continuous. Because of the truncation, we can only observe complete pairs when both T_{1i}, T_{2i} are greater than or equal to L_1, L_2 respectively.

Clayton (1978) introduced the function

$$\theta(t_1, t_2) = \frac{S(t_1, t_2)S^{(1,1)}(t_1, t_2)}{S^{(1,0)}(t_1, t_2)S^{(0,1)}(t_1, t_2)}$$

as a ratio of the hazard rate of conditional survivor function of T_1 , given $T_2 = t_2$, to that of T_1 , given $T_2 > t_2$. Oakes (1989) showed that $\theta(t_1, t_2) = \theta_0$ where θ_0 is a constant is a defining equation for the Clayton model. Thus, the Clayton survivor is

$$S(t_1, t_2) = \max \{S_1(t_1)^{1-\theta_0} + S_2(t_2)^{1-\theta_0} - 1, 0\}_+^{1/(1-\theta_0)}$$

where the subscript $+$ represents the positive part of the function is the unique solution to this equation when $\theta_0 \neq 1$. If $\theta_0 = 1$, T_1, T_2 are independent.

Kendall's tau, denoted by τ , is defined as:

$$\tau = E[\text{sign}\{T_{1i} - T_{1j}\}\text{sign}\{T_{2i} - T_{2j}\}],$$

where $(T_{1i}, T_{2i}), (T_{1j}, T_{2j})$ are independent copies of (T_1, T_2) . It is a dependence measure between random variables T_1 and T_2 and it relates with the unknown parameter θ through the equation:

$$\tau = 4 \int C_\theta(u_1, u_2) dC_\theta(u_1, u_2) - 1 = 4 \int S(t_1, t_2) dS(t_1, t_2) - 1.$$

For the Clayton model, $\tau = (\theta - 1)/(\theta + 1)$ for $\theta > 1$.

Now suppose that (T_{1i}, T_{2i}) ($i = 1, \dots, n$) follow the Clayton model and they are subject to left truncation (L_1, L_2) . Then the survivor function for truncated sample is $S(t_1, t_2)/S(L_1, L_2)$ for $t_1 \geq L_1, t_2 \geq L_2$ and

$$\begin{aligned} \theta(t_1, t_2) &= \frac{[S(t_1, t_2)/S(L_1, L_2)][S^{(1,1)}(t_1, t_2)/S(L_1, L_2)]}{[S^{(1,0)}(t_1, t_2)/S(L_1, L_2)][S^{(0,1)}(t_1, t_2)/S(L_1, L_2)]} \\ &= \frac{S(t_1, t_2)S^{(1,1)}(t_1, t_2)}{S^{(1,0)}(t_1, t_2)S^{(0,1)}(t_1, t_2)} = \theta_0 \end{aligned}$$

under the Clayton model assumption. Hence the truncated sample (with both components being larger than L_1 and L_2 respectively) should still follow the Clayton model with the same parameter value $\theta = \theta_0$. Conversely, assuming some absolutely continuous copula for (T_{1i}, T_{2i}) ($i = 1, \dots, n$), following the same arguments given in Oakes (2005), we can infer that the Clayton model assumption also holds for the original data (T_{1i}, T_{2i}) ($i = 1, \dots, n$). Based on this fact, we can set up goodness-of-fit test procedures for the original data by extending some existing test procedures for untruncated bivariate data. In this paper, we have developed our goodness-of-fit test by extending the test procedure for Archimedean copula models described in Genest, Quessy and Rémillard (2006a):

1. Let n_1 denote the number of pairs (T_{1i}, T_{2i}) such that $T_{1i} \geq L_1$ and $T_{2i} \geq L_2$. Define I_{ij} and $\hat{V}_i$ as:

$$I_{ij} = \begin{cases} 1 & \text{if } T_{1j} > T_{1i} \geq L_1 \text{ and } T_{2j} > T_{2i} \geq L_2 \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\hat{V}_i = \frac{\#\{(T_{1j}, T_{2j}) : T_{1j} > T_{1i} \geq L_1, T_{2j} > T_{2i} \geq L_2\}}{n_1 - 1} = \frac{\sum_{j \neq i} I_{ij}}{n_1 - 1}$$

respectively. For completely observed pairs (T_{1i}, T_{2i}) , applying the method proposed by Genest & Rivest (1993), we can estimate the unknown parameter by $\hat{\theta}_{n_1} = (1 + \hat{\tau}_{n_1})/(1 - \hat{\tau}_{n_1})$, where $\hat{\tau}_{n_1}$ is the analog estimator of τ , and it is defined as

$$\hat{\tau}_{n_1} = \frac{4}{n_1} \sum_{i=1}^{n_1} \hat{V}_i - 1.$$

2. Construct the test statistic based on the L^2 norm:

$$n_1 S(\theta) = n_1 \int_0^1 (\hat{K}_{n_1}(v) - K_\theta(v))^2 dv.$$

where $K_\theta(v) = \theta(1 - v^{\theta-1})/(\theta - 1)$ is the Kendall distribution function of the random variable $V_i = S(T_{1i}, T_{2i})$ for the Clayton model and $\hat{K}_{n_1}(v)$ is the corresponding empirical distribution function based on $\hat{V}_i$. Here we have applied the same notation as used in Wang & Wells (2000a). Basically we will replace $S(\theta)$ by:

$$S_1(\hat{\theta}_{n_1}) = \sum_{i=1}^{n_1} \{\hat{K}_{n_1}(v_{(i)}) - K_{\hat{\theta}_{n_1}}(v_{(i)})\}^2 (v_{(i)} - v_{(i-1)}),$$

where $v_0 = 0$ and $v_{(1)} \leq \dots \leq v_{(n_1)}$ are the ordered values of $\hat{V}_i$ $\{i = 1, 2, \dots, n_1\}$.

3. Treating $n_1 S_1(\hat{\theta}_{n_1})$ as the test statistic, we can obtain an empirical p-value in the following way:
 - Generate B random samples of size n_1 from the Clayton copula $C_{\hat{\theta}_{n_1}}$ and, for each of these samples, estimate θ by the same consistent parameter estimator and determine the value of the test statistic for this specific sample.

- If $S_{1:B}^* \leq \dots S_{B:B}^*$ denote the ordered values of the test statistics calculated in previous step, an estimate of the critical value of the test at level α based on $n_1 S_1(\hat{\theta}_{n_1})$ is given by $n_1 S_{[(1-\alpha)B]:B}^*$ and

$$\frac{1}{B} \# \{j : S_{j:B}^* \geq S_1(\hat{\theta}_{n_1})\}$$

yields an estimate of the p -value associated with the mean value $n_1 S_1(\hat{\theta}_{n_1})$ of the statistic.

The previous procedure was based on a general test procedure for Archimedean copula models proposed by Genest, Quessy and Rémillard (2006a). Their test procedure was initially developed for bivariate untruncated (uncensored) data and the effectiveness of the procedure was demonstrated by simulation studies in the same paper. In this paper, it has been extended to a test for the Clayton model for bivariate truncated data. The validity of the bootstrap procedure to calculate the p -value has been justified in Genest and Rémillard (2005). Here are the main steps they have taken to validate this procedure: our test statistic can be written as

$$n_1 S_1(\hat{\theta}_{n_1}) = n_1 \int_0^1 \{\hat{K}_{n_1}(v) - K_{\hat{\theta}_{n_1}}(v)\}^2 dv$$

where $\hat{K}_{n_1}(v) = \sum_{i=1}^{n_1} I(\tilde{V}_i \leq v)/n_1$ and $\hat{\theta}_{n_1}$ is the parameter estimator satisfying some regularity conditions. Bootstrap samples are then generated from $C_{\hat{\theta}_{n_1}}$ and

$$n_1 S_b^*(\hat{\theta}_{n_1}^*) = n_1 \int_0^1 \{\hat{K}_{n_1}^*(v) - K_{\hat{\theta}_{n_1}^*}(v)\}^2 dv$$

is calculated from each bootstrap sample. To prove that the bootstrap p -value is asymptotically correct, we only need to show that $n_1 S_1(\hat{\theta}_{n_1})$ and $n_1 S_b^*(\hat{\theta}_{n_1}^*)$ converges to the same distribution. This is equivalent to proving that the empirical process

$$n_1^{1/2} \{\hat{K}_{n_1}(v) - K_{\hat{\theta}_{n_1}}(v)\}$$

and the bootstrap empirical process

$$n_1^{1/2} \{\hat{K}_{n_1}^*(v) - K_{\hat{\theta}_{n_1}^*}(v)\}$$

converge to the same Gaussian process, where $\hat{K}_{n_1}^*(v)$ and $\hat{\theta}_{n_1}^*$ are estimated based on the bootstrap sample generated from $C_{\hat{\theta}_{n_1}}$. To this end, we further

decompose $n_1^{1/2}\{\hat{K}_{n_1}(v) - K_{\hat{\theta}_{n_1}}(v)\}$ into two terms:

$$\begin{aligned} & n_1^{1/2}\{\hat{K}_{n_1}(v) - K_{\hat{\theta}_{n_1}}(v)\} \\ &= n_1^{1/2}\{\hat{K}_{n_1}(v) - K_{\theta_0}(v)\} - n_1^{1/2}\{K_{\hat{\theta}_{n_1}}(v) - K_{\theta_0}(v)\} \end{aligned}$$

Barbe et al. (1996) have shown that the first empirical process on the right hand side converges to a Gaussian process K under C_{θ_0} (θ_0 is the true parameter value). Genest, Quessy and Rémillard (2006a) have shown that the second empirical process converges to $\dot{K}\Theta$ (where $\dot{K} = \partial K/\partial\theta$ and Θ is the limiting process of $n_1^{1/2}(\hat{\theta}_{n_1} - \theta)$) under C_{θ_0} . Therefore, $n_1^{1/2}\{\hat{K}_{n_1}(v) - K_{\hat{\theta}_{n_1}}(v)\}$ converges to a Gaussian process $K - \dot{K}\Theta$. Using LeCam's third lemma, one can show that

$$\begin{aligned} & n_1^{1/2}\{\hat{K}_{n_1}^*(v) - K_{\hat{\theta}_{n_1}^*}^*(v)\} \\ &= n_1^{1/2}\{\hat{K}_{n_1}^*(v) - K_{\theta_0}(v)\} - n_1^{1/2}\{K_{\hat{\theta}_{n_1}^*}^*(v) - K_{\theta_0}(v)\} \end{aligned}$$

converges to $(\tilde{K} + a\tilde{\Theta}) - \dot{K}(\tilde{\Theta} + b\tilde{\Theta})$ under $C_{\hat{\theta}_{n_1}}$, where $\tilde{K}$ and $\tilde{\Theta}$ are independent copies of K and Θ respectively, $a(v) = E(K(v)W^T)$, $b = E(\Theta W^T)$ and W is the score per observation. Notice the facts that our parameter estimator $\hat{\theta}_{n_1}$ and $\hat{K}_{n_1}(v)$ are all regular, as have been defined and proved in Genest and Rémillard (2005) (see section 4.2.2 and Proposition 4.3 in their paper). Therefore $a(v) = \dot{K}(v) = E(K(v)W^T)$ and $b = I$ (I is an identity matrix) respectively. From these facts we can conclude that the bootstrap empirical process

$$\begin{aligned} & n_1^{1/2}\{\hat{K}_{n_1}^*(v) - K_{\hat{\theta}_{n_1}^*}^*(v)\} \\ &= n_1^{1/2}\{\hat{K}_{n_1}^*(v) - K_{\theta_0}(v)\} - n_1^{1/2}\{K_{\hat{\theta}_{n_1}^*}^*(v) - K_{\theta_0}(v)\} \end{aligned}$$

converges to $(\tilde{K} + a\tilde{\Theta}) - \dot{K}(\tilde{\Theta} + I\tilde{\Theta}) = \tilde{K} - \dot{K}\tilde{\Theta}$ which is the same Gaussian process as $K - \dot{K}\Theta$. For more details, please see Genest and Rémillard (2005).

Alternatively, we can use Shih's test (1998) to test the Clayton model assumption for left truncated bivariate data. However, Shih's test is more complex than the one described here and according to Genest, Quessy and Rémillard (2006b), the original asymptotic variance of Shih's test statistic is erroneous. One should use the corrected Shih's test procedure described in Genest, Quessy and Rémillard (2006a and 2006b) to perform such a test.

3 Parameter Estimation

Once the previous test procedure justifies our Clayton model assumption, one can estimate the parameter value for original data based on truncated sample. Because the unknown parameter and Kendall's τ are linked by an one-to-one function $\tau = (\theta - 1)/(\theta + 1)$, we can choose Kendall's τ as the unknown parameter to estimate. As mentioned before, for the Clayton model, τ as a function of θ is unchanged under the truncation. Hence we can use truncated τ to estimate Kendall's τ for the original data. The reason why we choose truncated τ instead of the semi-parametric estimator proposed by Shih & Louis (1995) to estimate the unknown parameter for the Clayton model is that their procedure is based upon the log-likelihood function of the joint distribution. If the true distribution of original data is slightly different from the Clayton model while the p-value of our goodness-of-fit test is insignificant, the semi-parametric estimator may lose its efficiency. Compared with that, truncated τ is a more reasonable estimator as a correlation measure for two random variables since it is essentially nonparametric. Another advantage of truncated τ is its simplicity: there is no score equation to be solved. Under the Clayton model assumption, it is easy to see that the proposed estimate is still a consistent estimator for Kendall's τ .

Treating the truncated data as if it were complete, one can estimate the variance of the proposed estimator based on the following proposition:

Proposition 1 . *Under suitable regularity conditions, $n_1^{1/2}(\hat{\tau}_{n1} - \tau)$ is asymptotic normal with zero mean and variance*

$$\sigma^2 = 16E\left(V_1^2 + 2F(T_1, T_2)V_1 - 4V_1V_2 + P_{12}^{(1)}\right),$$

where $P_{12}^{(1)} = S\{\max(T_{11}, T_{12}, L_1), \max(T_{21}, T_{22}, L_2)\}$ and $F(t_1, t_2) = Pr(T_1 < t_1, T_2 < t_2)$.

The proof is given in the Appendix. Based on this Proposition, we can estimate the variance by corresponding sample estimates

$$\frac{16}{n_1^2} \sum_{i=1}^{n_1} \{\hat{V}_i^2 + 2\hat{F}(T_{1i}, T_{2i})\hat{V}_i\} - \frac{64}{n_1^3} \left\{ \sum_{i=1}^{n_1} \hat{V}_i \right\}^2 + \frac{16}{n_1^2(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} \hat{P}_{ij}^{(1)},$$

where $\hat{P}_{ij}^{(1)}$ equals to

$$\frac{\#\{(T_{1k}, T_{2k}) : T_{1k} > \max(T_{1i}, T_{1j}, L_1), T_{2k} > \max(T_{2i}, T_{2j}, L_2)\}}{n_1 - 2}$$

Under necessary regularity conditions, it is easy to show that the nonparametric variance estimate is asymptotically equivalent to

$$\frac{16}{n_1} E \left(V_1^2 + 2F(T_1, T_2)V_1 - 4V_1V_2 + P_{12}^{(1)} \right)$$

when n_1 gets large.

A simulation study was conducted to investigate the performance of truncated τ according to different levels of truncations. Paired data vectors whose marginal distributions are all unit-exponential are first generated from the Clayton model. We then generated truncation variables. For simplicity, we considered the truncation variables $L_1 = L_2 = 0.25, 0.50, 0.75, 1.00$ respectively and report the results for $n = 100$ and $n = 200$ in Tables 1 and 2. These truncation values correspond to light (about 20%) to heavy (about 75%) truncation. From Tables 1 and 2, we can see that under the Clayton model assumption, the parameter estimator $\hat{\tau}_{n_1}$ is rather accurate and our variance estimates based on Proposition 1 perform quite well when n is large.

Table 1: Performance of $\hat{\tau}_{n_1}$ for truncated data based on 1000 repetitions. For each τ value, the first row is the mean of the estimates, the second row is the mean of the sample variance estimates calculated based on our variance formula and the third row is the mean of the empirical variances.

Sample Size $n = 100$				
τ	$L_1 = L_2 = 0.25$	$L_1 = L_2 = 0.50$	$L_1 = L_2 = 0.75$	$L_1 = L_2 = 1.00$
0.2	0.1970	0.1979	0.1966	0.1957
	(0.0068)	(0.0105)	(0.0151)	(0.0216)
	(0.0075)	(0.0131)	(0.0189)	(0.0228)
0.4	0.3999	0.4026	0.4004	0.4017
	(0.0053)	(0.0074)	(0.0101)	(0.0131)
	(0.0057)	(0.0087)	(0.0113)	(0.0181)
0.6	0.5981	0.6011	0.5987	0.6019
	(0.0029)	(0.0039)	(0.0050)	(0.0061)
	(0.0032)	(0.0045)	(0.0064)	(0.0084)
0.8	0.7996	0.8016	0.8012	0.8010
	(0.0007)	(0.0008)	(0.0008)	(0.0007)
	(0.0012)	(0.0015)	(0.0021)	(0.0027)

Table 2: Performance of $\hat{\tau}_{n_1}$ for truncated data based on 1000 repetitions. For each τ value, the first row is the mean of the estimates, the second row is the mean of the sample variance estimates calculated based on our variance formula and the third row is the mean of the empirical variances.

Sample Size $n = 200$				
τ	$L_1 = L_2 = 0.25$	$L_1 = L_2 = 0.50$	$L_1 = L_2 = 0.75$	$L_1 = L_2 = 1.00$
0.2	0.2016	0.2042	0.1965	0.2046
	(0.0038)	(0.0063)	(0.0104)	(0.0176)
	(0.0040)	(0.0071)	(0.0121)	(0.0225)
0.4	0.4033	0.4004	0.3998	0.3968
	(0.0027)	(0.0039)	(0.0052)	(0.0070)
	(0.0028)	(0.0043)	(0.0054)	(0.0079)
0.6	0.5997	0.6002	0.5979	0.5967
	(0.0015)	(0.0020)	(0.0027)	(0.0034)
	(0.0015)	(0.0023)	(0.0029)	(0.0041)
0.8	0.7997	0.7994	0.8000	0.8007
	(0.0004)	(0.0005)	(0.0006)	(0.0007)
	(0.0005)	(0.0007)	(0.0009)	(0.0012)

It is worthy of pointing out that the proposed test and estimation procedures can also be applied to bivariate data under fixed left censoring, as will be shown in the next section. This is because we can form a subsample from such a data set by including only the completely observable pairs and treat it as if it were a truncated data set. Suppose that the original sample follows the Clayton model and the censoring vector is a fixed vector (L_1, L_2) , then the survivor function of completely observable pairs will still be $S(t_1, t_2)/S(L_1, L_2)$ (we have used the same notation as in section 2). Hence the arguments in Oakes (2005) can be reapplied here to show that the original uncensored data follows the Clayton model if and only if the completely observable pairs follows the same model. For details of the proof, please see Oakes (2005). Therefore, the same test and estimation procedure for bivariate truncated data can be applied for this type of bivariate censored data.

4 An Illustrative Example

A cohort study of HIV-infected men was conducted at the Hospital Univer-

sitrio Clementino Fraga Filho in Rio de Janeiro, Brazil to evaluate the effect of antiretroviral therapy on HIV shedding in semen (Barroso et al. 2000). One of the objectives of the cohort study was to assess the association between plasma and semen viral loads. HIV RNA was measured using the NucliSens quantification assay. The data set is presented in Tables 3 and 4. Of the 85 men who provided a blood sample and a semen sample for HIV RNA measurements, 75% (64) of the semen samples and 55% (47) of the blood samples have undetectable viral loads (detection limit is $L_1 = L_2 = 2.60$). Overall we have 19 men with complete observations. Since the detection limit is high, the bivariate data set is severely left censored at (L_1, L_2) . It is of scientific interest to find the original association between the plasma viral loads T_1 and the semen viral loads T_2 .

Table 3: HIV Data Set Part I

ID	1	2	3	4	5	6	7	8	9	10	11
$\min\{T_1, L_1\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6
$I\{T_1 > L_1\}$	0	0	0	0	0	0	0	0	0	0	0
$\min\{T_2, L_2\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	3.6	2.6	2.6
$I\{T_2 > L_2\}$	0	0	0	0	0	0	1	0	0	0	0
ID	12	13	14	15	16	17	18	19	20	21	22
$\min\{T_1, L_1\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6
$I\{T_1 > L_1\}$	0	0	0	0	0	0	0	0	0	0	0
$\min\{T_2, L_2\}$	2.6	2.6	2.6	3.8	2.6	2.6	2.6	2.6	3.2	2.6	3.0
$I\{T_2 > L_2\}$	0	0	0	1	0	0	0	0	1	0	1
ID	23	24	25	26	27	28	29	30	31	32	33
$\min\{T_1, L_1\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6
$I\{T_1 > L_1\}$	0	0	0	0	0	0	0	0	0	0	0
$\min\{T_2, L_2\}$	2.6	3.6	3.1	2.6	3.9	2.6	2.6	2.6	2.6	3.2	3.43
$I\{T_2 > L_2\}$	0	1	1	0	1	0	0	0	0	1	1
ID	34	35	36	37	38	39	40	41	42	43	44
$\min\{T_1, L_1\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6
$I\{T_1 > L_1\}$	0	0	0	0	0	0	0	0	0	0	0
$\min\{T_2, L_2\}$	3.18	2.6	2.6	2.6	4.11	2.6	2.6	2.6	2.6	3.9	3.7
$I\{T_2 > L_2\}$	1	0	0	0	1	0	0	0	0	1	1

Table 4: HIV Data Set Part II

ID	45	46	47	48	49	50	51	52	53	54	55
$\min\{T_1, L_1\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6
$I\{T_1 > L_1\}$	0	0	0	0	0	0	0	0	0	0	0
$\min\{T_2, L_2\}$	2.6	3.5	2.6	2.6	2.6	4.2	2.6	2.6	2.6	4.6	2.6
$I\{T_2 > L_2\}$	0	1	0	0	0	1	0	0	0	1	0
ID	56	57	58	59	60	61	62	63	64	65	66
$\min\{T_1, L_1\}$	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	3.0	3.0
$I\{T_1 > L_1\}$	0	0	0	0	0	0	0	0	0	1	1
$\min\{T_2, L_2\}$	2.6	2.6	3.3	2.6	2.6	2.6	2.8	4.8	2.6	3.3	4.9
$I\{T_2 > L_2\}$	0	0	1	0	0	0	1	1	0	1	1
ID	67	68	69	70	71	72	73	74	75	76	77
$\min\{T_1, L_1\}$	3.2	3.3	3.6	3.6	3.7	3.8	3.8	3.9	4.0	4.0	4.2
$I\{T_1 > L_1\}$	1	1	1	1	1	1	1	1	1	1	1
$\min\{T_2, L_2\}$	3.9	3.2	2.6	4.2	4.0	4.7	4.4	4.2	5.2	4.6	4.6
$I\{T_2 > L_2\}$	1	1	0	1	1	1	1	1	1	1	1
ID	78	79	80	81	82	83	84	85			
$\min\{T_1, L_1\}$	4.4	4.5	4.5	4.5	5.0	5.3	5.5	6.0			
$I\{T_1 > L_1\}$	1	1	1	1	1	1	1	1			
$\min\{T_2, L_2\}$	4.9	2.6	4.3	3.3	5.2	5.5	6.1	4.1			
$I\{T_2 > L_2\}$	1	0	1	1	1	1	1	1			

To this end, we first performed the proposed empirical test for the subsample formed by the completely observable pairs and obtained an empirical p-value 0.11. Hence there is not enough evidence to reject the null hypothesis that (T_1, T_2) follows the Clayton model. Under this model assumption, τ can be estimated by $\hat{\tau}_{n_1} = 4 \sum_{i=1}^{n_1} \hat{V}_i / n_1 - 1 = 0.37$ with the variance equal to 0.03 based on $n_1 = 19$ observed pairs. Hence we can conclude that there exists moderate association between the plasma viral loads T_1 and the semen viral loads T_2 .

5 Concluding Remarks

In this paper, we proposed an empirical test procedure to check the Clayton model assumption for bivariate truncated data. It is a generalization of the test procedure for Archimedean copula models described in Genest, Quessy and

Rémillard (2006a). From the simulation studies presented in Genest, Quessy and Rémillard (2006a), we can see that the proposed test procedure works quite well. Hence our test procedure can serve as an effective tool to check the goodness-of-fit for the Clayton model for bivariate data under left truncation or fixed left censoring.

Once the test result justifies the Clayton model assumption, truncated τ has been suggested as a simple estimator of the unknown parameter. From our simulation studies, we find that the estimator works quite well for heavily truncated bivariate data under the Clayton model assumption.

On the other hand, if our test result is against the Clayton model assumption, other nonparametric estimator of Kendall's τ such as Brown's estimator (1974) or Wang and Wells' estimator (2000b) can be applied. However, these estimators tend to be quite biased when the truncation levels are very high. This is because in general, the copula model structure is not preserved for other types of copula models under left truncation or fixed left censoring. Therefore, how to analyze such kind of truncated or censored data effectively is still an open question to the statistical society. Fortunately, many bivariate data can be modelled effectively by the Clayton model and the proposed strategies are expected to be useful for the correlation study in biomedical research.

Appendix

Proof of Proposition 1: First, notice the fact that $\hat{\tau}_{n1}$ is a U-statistic with the mean equal to τ . Hence $n_1^{1/2}(\hat{\tau}_{n1} - \tau)$ should be asymptotically normal with zero mean and some variance σ^2 according to the U-statistic theory. The asymptotic variance of $\hat{\tau}_{n1}$ can be derived in the following way:

$$\begin{aligned} \text{var}[\hat{\tau}_{n1}] &= \text{var}\left[4 \sum_{i=1}^{n_1} \hat{V}_i / n_1 - 1\right] = \frac{16}{n_1^2} \sum_{i=1}^{n_1} \text{var}(\hat{V}_i) + \frac{16}{n_1^2} \sum_{1 \leq i \neq j \leq n_1} \text{cov}(\hat{V}_i, \hat{V}_j) \\ &= \frac{16}{n_1^2} n_1 \{E[\text{var}(\hat{V}_i | V_i)] + \text{var}[E(\hat{V}_i | V_i)]\} + \frac{16n_1(n_1 - 1)}{n_1^2} \text{cov}(\hat{V}_i, \hat{V}_j). \end{aligned}$$

Given V_i , $(n_1 - 1)\hat{V}_i$ is distributed as Binomial($n_1 - 1, V_i$), one has $E(\hat{V}_i | V_i) = V_i$ and $\text{var}(\hat{V}_i | V_i) = V_i(1 - V_i)/(n_1 - 1)$. Suppose that

$$\text{cov}(\hat{V}_i, \hat{V}_j) = \frac{1}{(n_1)} E(P_{12}^{(1)} - 3V_1V_2 + I_{12}V_1 + I_{21}V_2) + o\left(\frac{1}{n_1}\right) \quad (A.1)$$

is true, we have:

$$\begin{aligned} \text{var}[\hat{\tau}_{n_1}] &= \frac{16}{n_1(n_1 - 1)} E\{V_1(1 - V_1)\} + \frac{16}{n_1} \text{var}(V_1) \\ &+ \frac{16n_1(n_1 - 1)}{n_1^2} \frac{1}{n_1} E(P_{12}^{(1)} - 3V_1V_2 + I_{12}V_2 + I_{21}V_1) + o\left(\frac{1}{n_1}\right) \\ &= \frac{16}{n_1} [E(V_1^2) - E^2(V_1)] + \frac{16}{n_1} [P_{12}^{(1)} - 3V_1V_2 + I_{12}V_2 + I_{21}V_1] + o\left(\frac{1}{n_1}\right). \end{aligned}$$

Because V_1 and V_2 are independent and identically distributed, we have $E^2(V_1) = E(V_1V_2)$. We also have

$$\begin{aligned} E(I_{12}V_2) &= E[E\{I_{12}V_2|(T_{12}, T_{22})\}] \\ &= E[V_2E\{I_{12}|(T_{12}, T_{22})\}] = E(S(T_{12}, T_{22})F(T_{12}, T_{22})). \end{aligned}$$

Similarly we have

$$E(I_{21}V_1) = E(S(T_{11}, T_{21})F(T_{11}, T_{21})) = E(S(T_{12}, T_{22})F(T_{12}, T_{22})) = E(I_{12}V_2).$$

Combining above results we can conclude that

$$\text{var}[\hat{\tau}_{n_1}] = 16E\left(V_1^2 + 2F(T_{11}, T_{21})V_1 - 4V_1V_2 + P_{12}^{(1)}\right) / n_1.$$

Now we only need to show (A.1). First we define N_{st} , $s, t \in \{1, 2\}$ as follows: conditioning on the values of the pairs (T_{11}, T_{21}) and (T_{12}, T_{22}) , clearly, the $(n_1 - 2)$ remaining sample points can then be allocated to the different cells of the following 2×2 table according to a multinomial distribution, namely:

	$I_{2m} = 1$	$I_{2m} = 0$
$I_{1m} = 1$	N_{11}	N_{12}
$I_{1m} = 0$	N_{21}	N_{22}

Here $m = 3, 4, \dots, n_1$ and N_{st} denotes the observed frequency associated with cell (s, t) , $s, t = 1, 2$. If $P_{(12)st}$ denotes the conditional probability associated with that cell given the 1st and 2nd pairs. It is clear the joint conditional distribution of the N_{st} is multinomial with parameters $(n_1 - 2)$ and $P_{(12)st}$, $s, t = 1, 2$. We also have: $(n_1 - 1)\hat{V}_1 = I_{12} + \sum_{m \neq 1, 2} I_{1m} = I_{12} + N_{11} + N_{12}$ and

$(n_1 - 1)\hat{V}_2 = I_{21} + \sum_{m \neq 1,2} I_{jm} = I_{21} + N_{11} + N_{21}$ by the definition of $\hat{V}_1$. The conditional expectation of $\{(n_1 - 1)\hat{V}_1\}\{(n_1 - 1)\hat{V}_2\}$ is given by

$$I_{12}I_{21} + (n_1 - 2)I_{12}V_2 + (n_1 - 2)I_{21}V_1 + (n_1 - 2)P_{12}^{(1)} \\ + (n_1 - 2)(n_1 - 3)(P_{12}^{(1)}P_{12}^{(1)} + P_{12}^{(1)}P_{(12)12} + P_{12}^{(1)}P_{(12)21} + P_{(12)12}P_{(12)21}).$$

Using the identities listed as follows: $P_{12}^{(1)} = S\{\max(T_{11}, T_{12}), \max(T_{21}, T_{22})\}$, $P_{(12)1+} = P_{12}^{(1)} + P_{(12)12} = S(T_{11}, T_{21}) = V_1$ and $P_{(12)+1} = P_{12}^{(1)} + P_{(12)21} = S(T_{22}, T_{22}) = V_2$. It is not hard to reach:

$$P_{12}^{(1)}P_{12}^{(1)} + P_{12}^{(1)}P_{(12)12} + P_{12}^{(1)}P_{(12)21} + P_{(12)12}P_{(12)21} = V_1V_2$$

We also know $E(\hat{V}|V) = V$. After some simple calculations, we can express the conditional covariance between $\hat{V}_1$ and $\hat{V}_2$ given (T_{11}, T_{21}) and (T_{12}, T_{22}) as:

$$\frac{1}{n_1} \left(P_{12}^{(1)} - 3V_1V_2 + I_{12}V_2 + I_{21}V_1 \right) + o_p \left(\frac{1}{n_1} \right).$$

Taking the expectation with respect to (T_{11}, T_{21}) and (T_{12}, T_{22}) , one obtain the desired result (A.1).

References

- Barbe, P., Genest, C., Ghouadi, K., and Remillard, B. (1996). On Kendall's process. *Journal of Multivariate Analysis* **58**, 197-229.
- Barroso, P. F., Schechter, M., Gupta, P., Melo, M. F., Vieira, M., Murta, F. C., Souza, Y., Harrizon, L. H., (2000). Effect of antiretroviral therapy on HIV shedding in semen. *Annals of Internal Medicine* **133**, 280-284.
- 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.
- 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.

- Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika* **74**, 549-555.
- Genest, C., Ghoudi, K. and Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**: 543-552.
- Genest, C. and Rivest, L. P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of American Statistical Association* **88**: 1034-1043.
- Genest, C., Quessy, J-F, Rémillard, B. (2006a). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics* **30**, 337-366.
- Genest, C., Quessy, J-F, Rémillard, B. (2006b). On the joint asymptotic behavior of two rank-based estimators of the association parameter in the gamma frailty model. *Statistics & Probability Letters* **76**, 10-18.
- Genest, C., Rémillard, B. (2005). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *GERAD* G-2005-51.
- Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73**, 671-678.
- Manatunga, A. K. and Oakes, D. (1999). Parametric analysis of matched pairs survival data. *Lifetime Data Analysis* **5**, 371 – 387.
- Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of American Statistical Association* **84**, 487-493.
- Oakes, D. (2005). On the preservation of the copula structure under truncation. *The Canadian Journal of Statistics* **97**, 77-87.
- Sklar, A. (1959), "Fonctions de rpartition n dimensions et leurs marges", Publications de l'Institut de Statistique de L'Universit de Paris 8, 229-231.

- Shih, J. and Louis, T. A. (1995). Inference on association parameter in copula models for bivariate survival data. *Biometrics* **26**, 183-214.
- Shih, J. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika* **85**, 189-200.
- Wang, W. and Wells, M. T. (2000a). Model selection and semiparametric inference for bivariate failure-time data. *Journal of American Statistical Association* **449**, 62-72.
- Wang, W. and Wells, M. T. (2000b). Estimation of Kendall's tau under censoring. *Statistica Sinica* **10**, 1199-1218.