Statistics and Its Interface Volume 11 (2018) 463–471

A new copula model-based method for regression analysis of dependent current status data

Qi Cui, Hui Zhao, and Jianguo Sun*

This paper discusses regression analysis of current status data, which arise when the occurrence of the failure event of interest is observed only once or the occurrence time is either left- or right-censored [5, 11]. Many authors have investigated the problem, however, most of the existing methods are parametric or apply only to limited situations such that the failure time and the observation time have to be independent. In particular, Ma et al. [7] recently proposed a copula-based procedure for the situation where the failure time and the observation time are allowed to be dependent but their association needs to be known. To address this restriction, we present a new two-step estimation procedure that allows one to estimate the association parameter in addition to estimation of other unknown parameters. The asymptotic properties of the resulting estimators are established and a simulation study is conducted and suggests that the proposed method performs well for practical situations. Also an illustrative example is provided.

Keywords and phrases: Copula model, Current status data, Informative censoring, Proportional hazards model.

1. INTRODUCTION

Current status data often arise in many fields including epidemiological studies, social studies and tumorigenicity experiments. In this situation, the failure time T of interest is observed only once at a censoring or observation time C. In other words, the failure time of interest is not exactly observed and either left- or right-censored. In addition, the failure time and the observation time may often be correlated and we usually refer such data as dependent current status data. One example of this latter case naturally occurs in the tumorigenicity experiments where the failure time of interest is the time to tumor onset. In these situations, current status data occur because the study animals are usually only observed at their death or sacrifice and one only knows the presence or absence of the tumor at the time. It is well-known that most of the tumors are between lethal and non-lethal and thus the tumor onset time and the death time tend to be correlated. In other words, one only observes dependent current status data for the tumor onset time.

*Corresponding author.

There exists an extensive literature on regression analysis of current status data but most of the existing methods are parametric or apply only to limited situations. It is well-known that parametric approaches are usually questionable unless the assumed model can be confirmed, and one situation that has been discussed by many authors is when the failure time and the observation time are independent completely or given covariates. As mentioned above, the independence assumption may not be valid in many situations. Several authors have recently considered regression analysis of dependent current status data, including Ma et al. [7] and Zhao et al. [14] who proposed some copula model-based estimation procedures. The former studied the case where the failure time of interest marginally follows the proportional hazards model, while the latter discussed the case where the failure time marginally follows the additive hazards model. However, both methods assumed that the association parameter between the failure time of interest and the observation time is known, which is clearly not realistic in general.

As pointed out by [7] and others, the resulting estimators of regression parameters can be sensitive to the assumed association parameter or be biased and yield misleading results if the assumed association is misspecified. In the following, we will present a copula model-based estimation procedure that does not require the assumption. Some discussion on why this is possible will be given below. The copula model-based approach is a commonly used method for modeling correlated random variables or the association. For example, Shih and Louis [10] developed such estimation procedures for estimation of the association parameter based on bivariate right-censored data, and Wang et al. [13] generalized the method to the case of bivariate current status data. For the proposed method given below, as all authors mentioned above, we will assume that the copula model is known and some discussion on this will be given below.

The remainder of this paper is organized as follows. In Section 2, we will first introduce some notation and models to be used and present the resulting likelihood function. In particular, the copula model will be used to describe the association between the failure time and the observation time. Section 3 will describe the proposed sieve maximum likelihood estimation approach, a two-step estimation procedure. In addition, the resulting estimators of regression parameters will be shown to be consistent and asymptotically follow

a normal distribution. In Section 4, we will present some results obtained from an extensive simulation study conducted to assess the finite sample performance of the proposed approach, which indicate that the method seems to work well for practical situations. Section 5 applies it to a tumorigenicity study that motivated this study and Section 6 contains some discussions and concluding remarks.

2. NOTATION, ASSUMPTIONS AND THE LIKELIHOOD FUNCTION

Consider a failure time study that involves n independent subjects and in which each subject is observed only once. For subject i, let T_i denote the failure time of interest and Z_i a p-dimensional vector of covariates, and suppose that there exist two potential observation or censoring times denoted by C_i and ζ_i, $i = 1, ..., n$. Here we assume that C_i may be related to T_i but ζ_i is independent of T_i such as the administrative stop time. In the tumor example, C_i denotes the death time and ζ_i represents the sacrifice or study stopping time. Define $\tilde{C}_i = \min(C_i, \zeta_i)$, $\Delta_i = I(C_i \le \zeta_i)$ and $\delta_i = I(T_i \le \tilde{C}_i)$. Then the observed data have the form $\{ X_i = (\Delta_i, \delta_i, \tilde{C}_i, Z_i),\ i = 1, \ldots, n \}$.

To describe the effects of covariates, in the following, we will assume that given the covariates Z_i's, T_i and C_i follow the marginal proportional hazards models given by

$$\lambda^{(T)}(t \mid Z_i) = \lambda_1(t) \exp(Z_i^T \beta) \tag{1}$$

and

$$\lambda^{(C)}(t \mid Z_i) = \lambda_2(c) \exp(Z_i^T \gamma)\,, \tag{2}$$

respectively. Let F_T and F_C denote the marginal distributions of the T_i's and the C_i's given covariates, respectively, and F the joint distribution of T_i and C_i. Then there exists a copula function $C_\alpha(u, v)$ defined on $I^2 = [0, 1] \times [0, 1]$ such that

$$F(t, c) = C_\alpha\{F_T(t), F_C(c)\} \tag{3}$$

[8]. Here α is often referred to as association parameter representing the relationship between T_i and C_i, and $C_\alpha(u, 0) = C_\alpha(0, v) = 0$, $C_\alpha(u, 1) = u$ and $C_\alpha(1, v) = v$. It follows that

$$\begin{aligned} P(T \le t \mid C = c, Z_i) &= \frac{\partial C_\alpha(u, v)}{\partial v} \mid_{u=F_T(t), v=F_C(c)} \\ &= m_\alpha\{F_T(t), F_C(c)\}\,. \end{aligned}$$

Note that the copula model-based approach is one of the most commonly used methods for modeling or dealing with correlated random variables and this is especially the case for correlated failure time variables in either bivariate or multivariate failure time data analysis [2, 10, 13].

Define $\Lambda_T(t) = \int_0^t \lambda_1(s)ds$ and $\Lambda_C(c) = \int_0^t \lambda_2(s)ds$, and let f_C denote the marginal density function of the C_i's given covariates. Then under the assumptions above, we have

$$\begin{aligned} F_T(t) &= 1 - \exp\{-\Lambda_T(t)\, e^{Z_i^T \beta}\}, \\ F_C(c) &= 1 - \exp\{-\Lambda_C(c)\, e^{Z_i^T \gamma}\}, \end{aligned}$$

and

$$f_C(c) = \exp\{-\Lambda_C(c)\, e^{Z_i^T \gamma}\}\, \lambda_2(c)\, \exp(Z_i^T \gamma)\,.$$

Furthermore the resulting likelihood function can be written as

$$\begin{aligned} L(\theta, \eta) = \prod_{i=1}^{n} \Big\{ & \Big([m_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{\delta_i} \\ & [1 - m_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{1-\delta_i} f_C(\tilde{C}_i) \Big)^{\Delta_i} \\ & \times \Big([F_T(\tilde{C}_i) - C_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{\delta_i} \\ & [1 - F_T(\tilde{C}_i) - F_C(\tilde{C}_i) \\ & + C_\alpha\{F_T(\tilde{C}_i), F_C(\tilde{C}_i)\}]^{1-\delta_i} \Big)^{1-\Delta_i} \Big\}, \end{aligned}$$

where $\theta = \{\beta^T, \alpha, \Lambda_T(\cdot)\}^T$ and $\eta = \{\gamma^T, \Lambda_C(\cdot)\}^T$. In the next section, we will discuss estimation of regression parameters as well as other parameters.

3. INFERENCE PROCEDURE

Now we will discuss the estimation and inference about models 1 and 2 with the focus on regression parameter β. For this, we will present a two-step sieve estimation procedure that first estimates model 2 and then model 1. More specifically, for the first step, note that for the observation time C_i's, we have complete or right-censored data and thus it is natural to estimate γ and Λ_C by the maximum partial likelihood estimator and Breslow estimator, respectively.

3.1 Two-step sieve estimation procedure

Let $\hat{\gamma}$ and $\hat{\Lambda}_C$ denote the estimators of γ and Λ_C defined above, respectively. Then one can estimate the marginal distribution of the C_i's by $\hat{F}_C(c) = 1 - \exp\{-\hat{\Lambda}_C(c) \exp(Z_i^T \hat{\gamma})\}$. Given $\hat{\eta} = (\hat{\gamma}, \hat{\Lambda}_C)$, for the second step, to estimate θ, it is apparent that one could maximize the conditional likelihood function

$$\begin{aligned} L(\theta|\hat{\eta}) = \prod_{i=1}^{n} \Big\{ & \Big([m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{\delta_i} \\ & [1 - m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{1-\delta_i} \hat{f}_C(\tilde{C}_i) \Big)^{\Delta_i} \\ & \times \Big([F_T(\tilde{C}_i) - C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{\delta_i} \\ & [1 - F_T(\tilde{C}_i) - \hat{F}_C(\tilde{C}_i) \\ & + C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}]^{1-\delta_i} \Big)^{1-\Delta_i} \Big\}. \end{aligned}$$

On the other hand, it is easy to see that this maximization can be difficult due to the dimension of $\Lambda_T(\cdot)$. To address this, by following Huang and Rossini [4] and others, we propose to approximate $\Lambda_T(\cdot)$ with monotone cubic I-splines first before the maximization [6, 9].

More specifically, let M denote a positive constant and $\{I_j(t)\}_{j=1}^{m+k_n}$ the I-spline base functions with order m and k_n interior knots, where $k_n = o(n^v)$ with $0 < v < 0.5$. The selection of m and k_n will be discussed below. Define

$$\Theta_n = \{\theta_n = (\beta^T, \alpha, \Lambda_{T_n})^T\} = \mathcal{B} \otimes \mathcal{M}_n\,,$$

where $\mathcal{B} = \{(\beta^T, \alpha)^T \in R^{p+1}, \| \beta \| + \| \alpha \| \leq M\}$ with $\| v \|$ denoting the Euclidean norm for a vector v, $\mathcal{M}_n = \{\Lambda_{T_n} : \Lambda_{T_n}(t) = \sum_{j=1}^{m+k_n} \xi_j I_j(t), \xi_j \geq 0, j = 1, \ldots, m + k_n, t \in [0, u_c]\}$, with u_c being the upper bound of all observation times $\{\tilde{C}_i : i = 1, \ldots, n\}$. It follows from Lemma A_1 of [6] that Θ_n can be used as a sieve space of the original parameter space Θ. Then we can estimate θ or $\theta = (\beta^T, \alpha, \xi^T)^T$ by the sieve maximum likelihood estimator, denoted by $(\hat{\beta}, \hat{\alpha}, \hat{\Lambda}_T(\cdot))$, defined as the value of θ that maximize the conditional log-likelihood function $l(\theta|\hat{\eta}) = \sum_{i=1}^n l^{(i)}(\theta|\hat{\eta})$, where $\xi^T = (\xi_1, \ldots, \xi_{m+k_n})$ and

$$\begin{aligned} l^{(i)}(\theta|\hat{\eta}) = {} & \Delta_i \log \hat{f}_C(\tilde{C}_i) \\ & + (1 - \delta_i)\Delta_i \log \left[1 - m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}\right] \\ & + \delta_i \Delta_i \log \left[m_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}\right] \\ & + \delta_i(1 - \Delta_i) \log \left[F_T(\tilde{C}_i) - C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}\right] \\ & + (1 - \delta_i)(1 - \Delta_i) \log \left[1 - F_T(\tilde{C}_i) - \hat{F}_C(\tilde{C}_i) \right. \\ & \left. + C_\alpha\{F_T(\tilde{C}_i), \hat{F}_C(\tilde{C}_i)\}\right]. \end{aligned}$$

As mentioned above, a main advantage of the estimation procedure proposed above over that given in [7] is that the former does not require that the association parameter is known. Also as discussed above, in general, the copula model and association parameter cannot be estimated without extra information. For the situation here, the extra information is given by the estimation of the marginal distribution F_C in the first step, which can then be treated as being known. The estimators $\hat{\gamma}$ and $\hat{\Lambda}_C$ have been studied by many authors and in particular, they are consistent [5]. In the next subsection, we will establish the asymptotic properties of $\hat{\beta}$ and discuss some implementation issues.

3.2 Asymptotic properties and implementation

Now we will establish the asymptotic properties of $\hat{\beta}$ and then discuss the variance estimation and some implementation issues.

Theorem 3.1. *Assume that the regularity conditions (C1)–(C4) described in the Appendix A and the conditions required in Lemma A.1 given in the Appendix A hold. Then as $n \to \infty$, we have that $\hat{\beta}$ is consistent and $\sqrt{n}\,(\hat{\beta} - \beta_0)$ converges to the multivariate normal distribution with mean zero, where β_0 denotes the true value of β.*

The proof of the results given above is sketched in the Appendix A. For the estimation of the covariance matrix of $\hat{\beta}$, one natural way would be to derive a consistent estimator but as can be seen in the Appendix A, such an estimator would be too complicated to be useful. Thus instead we suggest to employ the following bootstrap procedure. Let B denote a prespecified positive integer and for each $b = 1, \ldots, B$, draw a simple random sample $\{X_i^{(b)} = (\Delta_i^{(b)}, \delta_i^{(b)}, \tilde{C}_i^{(b)}, Z_i^{(b)}), i = 1, \ldots, n\}$ of size n with replacement from the observed data $\{X_i, i = 1, \ldots, n\}$. Let $\hat{\beta}^{(b)}$ denote the sieve maximum likelihood estimator of β defined above based on the resampled data set $\{X_i^{(b)}, i = 1, \ldots, n, b = 1, \ldots, B\}$. Then one can estimate the covariance matrix of $\hat{\beta}$ by

$$\hat{Var}(\hat{\beta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}^{(b)} - \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^{(b)}\right)^2.$$

Similarly one can show that $\hat{\alpha}$ is consistent and asymptotically follows a normal distribution too and estimate its variance by using the same approach.

To implement the estimation procedure above, it is apparent that one needs to specify m and k_n. In general, the degree m should be decided by the smoothness of the true baseline cumulative hazard function, and usually either quadratic or cubic spline functions work sufficiently well. Of course, one could try different values of them and compare the obtained results. As an alternative, one can apply the AIC to choose m and k_n that give the smallest AIC. Two other choices that one needs to make for the implementation of the estimation method proposed above are the interior knots and the copula model. For the former, a common way is to use the equally spaced quantiles for given m and k_n as discussed in the numerical studies below. For the latter, one may want to try some commonly used copula models and compare the obtained results, or also employ the AIC to choose it along with m and k_n together. The simulation studies below indicate that the proposed estimators seem to be relatively robust with respect to the copula model.

Also given m and k_n, the computation of $\hat{\theta}$ is relatively easy as one can simply employ some existing software such as the R function *nlminb*. In this, one does need to pay attention to the non-negativity constraint on the I-spline coefficients or as an alternative, one can avoid it by applying the logarithm transformation of the original coefficients before applying *nlminb*. For the selection of B in the variance estimation, one may start with some reasonable number and then increase it until the obtained results are stable. For a simulation study with large replications, one may choose small numbers such as $B = 50$ to save the computational effort.

Table 1. Estimation of the association and regression parameters under the Clayton model

Parameter	True	Bias	SSE	SEE	CP	True	Bias	SSE	SEE	CP
τ	0.2	0.036	0.237	0.240	0.961	0.2	0.040	0.219	0.221	0.960
β	0.4	-0.012	0.222	0.231	0.954	0.8	0.016	0.220	0.232	0.965
γ	0	0.003	0.142	0.145	0.943	0.4	-0.004	0.143	0.147	0.952
τ	0.2	0.050	0.248	0.238	0.931	0.2	0.050	0.222	0.223	0.948
β	0.4	0.023	0.212	0.220	0.960	0.8	0.029	0.242	0.249	0.954
γ	0.4	0.004	0.148	0.147	0.952	0.8	0.011	0.151	0.153	0.953
τ	0.4	-0.036	0.213	0.209	0.948	0.4	-0.039	0.194	0.197	0.967
β	0.4	0.018	0.208	0.221	0.965	0.8	0.003	0.205	0.217	0.964
γ	0	0.002	0.146	0.145	0.952	0.4	-0.006	0.147	0.147	0.955
τ	0.4	-0.047	0.221	0.210	0.956	0.4	-0.030	0.197	0.194	0.964
β	0.4	-0.011	0.210	0.207	0.958	0.8	-0.021	0.229	0.232	0.957
γ	0.4	-0.003	0.145	0.147	0.952	0.8	0.000	0.155	0.153	0.952

4. A SIMULATION STUDY

An extensive simulation study was conducted to evaluate the finite sample performance of the two-step estimation procedure proposed in the previous sections. In the study, we considered two situations for covariates and one is that there exists only one covariate generating from the Bernoulli distribution with the success probability of 0.5. The other is that there exist two covariates with one following the Bernoulli distribution with the success probability of 0.5 and the other following the uniform distribution over $(0, 1)$. To generate the failure times T_i's and the observation times C_i's, we took $\lambda_1(t) = \lambda_2(c) = 1$ in models (1) and (2) and first generated two independent random numbers u_i and w_i from the uniform distribution over $(0, 1)$. Then after obtaining the random number v_i by solving the equation

$$P(C \le c_i \mid T = t_i, Z) = \frac{\partial C_\alpha(u, v)}{\partial u} \mid_{u=u_i, v=v_i} = w_i \,,$$

we define $T_i = t_i$ and $C_i = c_i$, where t_i and c_i denote the solutions to the equations $F_T(t_i) = u_i$ and $F_C(c_i) = v_i$, respectively. The independent observation times ζ_i's were taken to be a constant.

For the generation of the T_i's and C_i's, we considered several copula models including the Clayton and Gumbel models given by

$$C_\alpha(u, v) = (u^{-(\alpha-1)} + v^{-(\alpha-1)} - 1)^{-1/(\alpha-1)}, \quad \alpha > 1 \,,$$

and

$$C_\alpha(u, v) = \exp[-\{(-\log u)^\alpha + (-\log v)^\alpha\}^{1/\alpha}], \quad \alpha \ge 1 \,.$$

Note that for different copula models, the spaces of the association parameter α are quite different and their interpretation also differs from case to case. Thus in the following, as others, we instead used the Kendall's τ, also a global association parameter, defined as

$$\tau = P\{(T_i - T_j)(C_i - C_j) > 0\} - P\{(T_i - T_j)(C_i - C_j) < 0\}$$

for i.i.d replicates (T_i, C_i) and (T_j, C_j) of (T, C). The Kendall's τ is always between -1 and 1 with 0 indicating independent and it is usually more robust and invariant to monotone transformation. The results below are based on 1,000 replications with the sample size $n = 200$.

Table 1 presents the results obtained on estimation of regression parameters β and γ as well as the Kendall's τ based on the simulated data generated under the Clayton model with one covariate and different true values for β, γ and τ. The results include the estimated bias given by the average of the proposed estimates minus the true value (Bias), the sample standard deviation (SSE), the average of estimated standard errors (SEE), and the 95% empirical coverage probability (CP). Here for the approximation of $\Lambda_T(\cdot)$, we used quadratic splines with the 0.2, 0.4, 0.6 and 0.8 quantiles of the $\tilde{C}_i$'s as four interior knots. The results obtained under the Gumbel model are given in Table 2 with two covariates and the other set-ups being the same as with Table 1. One can see from Table 1 and Table 2 that on the regression parameters, the proposed estimator seems to be unbiased and the bootstrap variance estimation also appears to be reasonable. In addition, the empirical coverage probabilities indicate that the normal approximation to the distributions of the proposed estimators seems to be appropriate. On the association parameter or Kendall's τ, it is clear that the performance of the proposed method is not as good as for the regression parameters and we believe that this is mainly because the estimated association parameter has a slower convergence than the estimated regression parameters.

As mentioned above, Ma et al. [7] discussed the same problem but their method assumes that the association parameter α or Kendall's τ is known. Thus it would be interesting to compare the proposed method to that given in [7]. For this, we repeated the studies above by applying both methods to the simulated data and present the estimation results in Table 3 with the true value of τ being 0.4. They suggest that both approaches performed well and the proposed method gave similar results to those given by [7]. The

Table 2. Estimation of the association and regression parameters under the Gumbel model

Parameter	True	Bias	SSE	SEE	CP	True	Bias	SSE	SEE	CP
τ	0.3	0.083	0.248	0.233	0.903	0.3	0.084	0.249	0.236	0.902
β_1	0.3	-0.004	0.194	0.207	0.968	0.3	-0.005	0.201	0.208	0.965
β_2	-0.3	0.015	0.336	0.351	0.960	0.3	0.003	0.333	0.351	0.974
γ_1	0.3	-0.000	0.140	0.147	0.960	0.3	0.005	0.151	0.147	0.950
γ_2	-0.3	-0.000	0.268	0.255	0.936	0.3	-0.001	0.253	0.255	0.951
τ	0.3	0.073	0.233	0.221	0.904	0.3	0.077	0.229	0.225	0.927
β_1	0.6	-0.010	0.223	0.226	0.961	0.6	-0.003	0.226	0.226	0.951
β_2	-0.6	-0.001	0.335	0.360	0.962	0.6	0.007	0.341	0.359	0.958
γ_1	0.6	0.004	0.149	0.150	0.955	0.6	0.007	0.153	0.150	0.949
γ_2	-0.6	-0.009	0.258	0.258	0.952	0.6	0.008	0.255	0.257	0.952
τ	0.6	-0.055	0.201	0.192	0.922	0.6	-0.043	0.200	0.191	0.922
β_1	0.3	-0.005	0.195	0.197	0.959	0.3	-0.008	0.183	0.194	0.958
β_2	-0.3	-0.007	0.323	0.331	0.965	0.3	-0.014	0.306	0.330	0.969
γ_1	0.3	0.012	0.151	0.147	0.943	0.3	0.007	0.150	0.147	0.946
γ_2	-0.3	-0.015	0.261	0.255	0.949	0.3	0.003	0.257	0.255	0.952
τ	0.6	-0.050	0.199	0.185	0.898	0.6	-0.049	0.199	0.186	0.917
β_1	0.6	-0.029	0.199	0.211	0.961	0.6	-0.023	0.203	0.210	0.939
β_2	-0.6	0.011	0.323	0.337	0.964	0.6	-0.018	0.329	0.340	0.956
γ_1	0.6	0.000	0.150	0.150	0.951	0.6	0.002	0.155	0.150	0.944
γ_2	-0.6	-0.011	0.260	0.257	0.948	0.6	0.013	0.259	0.257	0.947

Table 3. Comparison of the proposed method and the method given in Ma et al. (2015)

		Proposed method				Ma et al. (2015)			
Parameter	True	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
				Under Clayton model					
β	0.4	0.018	0.208	0.221	0.965	0.013	0.183	0.190	0.952
γ	0	0.002	0.146	0.145	0.952	0.005	0.145	0.145	0.943
β	0.4	-0.011	0.210	0.207	0.958	0.015	0.193	0.188	0.941
γ	0.4	-0.003	0.145	0.147	0.952	0.003	0.143	0.145	0.945
β	0.8	0.003	0.205	0.217	0.964	0.021	0.193	0.197	0.958
γ	0.4	-0.006	0.147	0.147	0.955	0.001	0.145	0.146	0.952
β	0.8	-0.021	0.229	0.232	0.957	0.024	0.210	0.205	0.941
γ	0.8	0.000	0.155	0.153	0.952	0.013	0.151	0.152	0.944
				Under Gumbel model					
β_1	0.4	-0.014	0.196	0.205	0.960	0.002	0.186	0.190	0.951
β_2	0	0.003	0.327	0.341	0.953	0.011	0.324	0.325	0.951
γ_1	0.4	-0.000	0.151	0.148	0.950	0.003	0.151	0.150	0.941
γ_2	0	-0.002	0.257	0.254	0.942	-0.000	0.253	0.257	0.946
β_1	0.4	-0.010	0.198	0.207	0.958	0.012	0.185	0.191	0.953
β_2	0.4	-0.022	0.341	0.346	0.960	-0.002	0.332	0.329	0.941
γ_1	0.4	0.005	0.150	0.148	0.946	0.008	0.148	0.149	0.952
γ_2	0.4	-0.014	0.251	0.256	0.961	-0.011	0.247	0.258	0.959

only difference is that as expected, the method given by [7] seems to be little more efficient than the proposed one on estimation of β, but they had similar efficiency on estimation of γ. We also considered other copula models including the FGM and Frank models described in the next section and obtained similar results.

5. AN APPLICATION

Now we apply the estimation approach proposed in the previous sections to a tumorigenicity study conducted by the National Toxicology Program and discussed in [7] among others. It is a 2-year study and consists of the groups of 50 male and 50 female F344/N rats and B6C3F$_1$ mice exposed to chloroprene at different concentrations. During the study, some animals died naturally during the study, and those who were alive at the end of study were sacrificed for the examination. Since the tumor status was only examined at the death or the end, we only have current status data for the tumor onset time, and one major goal of the study is to compare the tumor growth rates between the different dose

groups. Following [7], in the analysis below, we will focus on the specific type of lung tumor, the alveolar/bronchiolar adenoma, for the B6C3F$_1$ mice in the control group with no chloroprene inhalation and the high dose group with 80 ppm chloroprene inhalation with 100 in each of the two groups.

For the ith animal, let T_i denote the tumor onset time and C_i the death time with ζ_i representing the end of the study. Also define $Z_i = 1$ if the ith mice was in the high dose group and $Z_i = 0$ otherwise. For the analysis, we first considered the possible group effect on the death time and the application of the first step of the proposed estimation procedure yielded $\hat{\gamma} = 1.384$ with the estimated standard error of 0.19. This suggests that the animals in the high dose group had significantly a higher death rate than those in the control group. For the second step of the proposed estimation procedure, in addition to the Clayton and Gumbel models, we also considered a few other commonly used copula models including the FGM and Frank models defined, respectively, by

$$C_\alpha(u,v) = uv + \alpha uv(1-u)(1-v), \quad -1 \le \alpha \le 1\,,$$

and

$$C_\alpha(u,v) = -\frac{1}{\alpha}\log\Big\{1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\Big\}, \quad \alpha \ne 0\,.$$

Table 4 gives the obtained results on the estimation of the Kendall's τ and regression parameter β under the FGM and Gumbel models and they include the estimated parameter, the estimated standard errors and the p-values for testing the parameter equal to zero. Note that the results obtained under the Clayton and Frank models are similar to these under the Gumbel model and thus were not presented. Actually for the analysis here, all of the copula models considered gave consistent results. Here for the approximation to the baseline cumulative function $\Lambda_T(t)$, we employed the quadratic splines with $k_n = 3$, 4 or 5 and interior knots chosen the same way as in the simulation study above. Also we calculated and used the AIC for the selection of the appropriate copula model and k_n and found that the smallest AIC was given by the FGM copula with $k_n = 3$.

One can see from Table 4 that with respect to the dose effect, all results indicated that there existed a significant difference between the tumor growth rates of the animals in the two groups. The animals in the high dose group seem to have a significantly higher chance of developing the tumor than in the control group. With respect to estimation of the association between the tumor onset time and the death time, under the FGM model, the results suggest that they were significantly positively correlated. However, under the Gumbel model, although they seem to be positively correlated, the association level was not significant. In comparison, [7] also considered several copula models and employed the AIC to select appropriate copula model and the degree of the association. They also concluded that the FGM model provided the best fit and there was some significant dose effect on the tumor growth rate.

Table 4. Analysis results of the tumorigenicity experiment under FGM and Gumbel models

	Parameter	Estimate	SEE	p-value	AIC
		Under FGM model			
$k_n = 3$	τ	0.1680	0.0793	0.0342	721.1570
	β	2.4459	0.3880	<0.0001	
$k_n = 4$	τ	0.2222	0.0797	0.0053	722.2194
	β	2.5291	0.4069	<0.0001	
$k_n = 5$	τ	0.2222	0.0710	0.0017	724.8213
	β	2.4835	0.4050	<0.0001	
		Under Gumbel model			
$k_n = 3$	τ	0.0864	0.1610	0.5912	722.5539
	β	2.4514	0.4386	<0.0001	
$k_n = 4$	τ	0.0594	0.1766	0.7364	724.1031
	β	2.4659	0.4651	<0.0001	
$k_n = 5$	τ	0.0252	0.1881	0.8930	726.1204
	β	2.3306	0.4491	<0.0001	

6. DISCUSSIONS AND CONCLUSION REMARKS

This paper discussed regression analysis of current status data in the presence of dependent censoring or observation process and as mentioned above, such data occur quite often in many fields. This is in particular the case in tumorigenicity experiments where one has to deal with them almost always and in which an extensive literature, mainly parametric approaches, has been developed. For the problem, we presented a two-step copula model-based approach that allows one to estimate the association parameter in addition to regression parameters. It can be regarded as a generalization of the method given in [7], which assumes that the association level is known. The resulting estimators of regression parameters are consistent and their distributions can be asymptotically approximated by the normal distribution. Also the simulation study suggests that the proposed method seems to work well for practical situations.

Note that in the proposed method, we have employed I-spline functions to approximate the baseline cumulative hazard function Λ_T. As an alternative, one can use other smooth functions such as kernel functions and the method can be developed similarly. Also instead of using the sieve approach or approximating Λ_T by using smooth functions, one may directly maximize the conditional likelihood function $L(\theta|\hat{\eta})$ or employ the nonparametric maximum likelihood estimation. One main advantage of the sieve approach over the latter is that the maximization and its implementation can be much simpler and in the meantime, the two approaches can be asymptotically equivalent [4, 7].

One limitation of the proposed method is that it assumes that the underlying copula model is known. However, as many papers pointed out [7, 15], it is usually difficult or impossible to estimate it without strong assumptions. Also the simulation study suggested that the presented method seems to be robust with respect to the underlying copula

model. Another limitation is that it has been assumed that both the failure time of interest and the informative censoring or observation time follow the proportional hazards model and it is apparent that this may not hold in practice. For example, one of them or both may follow other models such as the additive hazards model or linear transformation model. It is straightforward to generalize the proposed estimation procedure to these situations. Also in the preceding sections, we have focused only on current status data, a special case of interval-censored data, and the similar problem can often occur for general interval-censored data too [11]. For the latter case, although the idea described above can still be applied, the development would be much more complicated and difficult partly as one may have to deal with three dimensional copula models.

APPENDIX A. PROOF OF THE ASYMPTOTIC PROPERTIES

For the proof and the completeness, we first describe the asymptotic properties of $\hat{\gamma}$ and $\hat{\Lambda}_C$ in the Lemma A.1 below.

Lemma A.1. *Let $\hat{\gamma}$ and $\hat{\Lambda}_C$ be the estimators of γ and Λ_C defined above, respectively, and assume that the regularity conditions given at pages 174–176 of [5] hold. Then $\hat{\gamma}$ and $\hat{\Lambda}_C$ are consistent and have the asymptotical normality.*

For the proof of the results above, the readers are referred to Theorems 8.3.1, 8.3.2 and 8.3.3 of [1]. To show the asymptotic properties of $\hat{\beta}$, in addition to the conditions needed in Lemma A.1, we also need the following regularity conditions.

(C1) For the follow-up time τ, we have $P(\tau \geq \tau_0) > 0$;

(C2) The covariate Z has a bounded support in R^p.

(C3) (i) The copula function $C(\cdot,\cdot)$ has bounded first order partial derivatives with $\partial C(u,v)/\partial u$ and $\partial C(u,v)/\partial v$ being Lipschitz. (ii). Assume that $\mu_C(E) > 0$ for any open set $E \in I^2$, where μ_C denotes the probability measure corresponding to the copula function C given Z.

(C4) The κth derivative of $\Lambda_T(\cdot)$, denoted by $\Lambda_T^{(\kappa)}(\cdot)$, is Holder continuous such that $|\Lambda_T^{(\kappa)}(t_1) - \Lambda_T^{(\kappa)}(t_2)| \leq M_0|t_1 - t_2|^\eta$ for some $\eta \in (0,1]$ and any $t_1, t_2 \in (0, u_c)$, where M_0 is a constant.

Proof of the consistency of $\hat{\theta}$. Let $\theta_0 = (\beta_0^T, \alpha_0, \Lambda_{T0})^T$ and $\eta_0 = (\gamma_0^T, \Lambda_{C0})^T$ denote the true values of θ and η, respectively. For any $\epsilon > 0$, based on the fact that $\{|l(\hat{\theta}|\hat{\eta}) - l(\theta_0|\eta_0)| \geq \epsilon\} \subset \{|l(\hat{\theta}|\hat{\eta}) - l(\hat{\theta}|\eta_0)| \geq \epsilon/2\} \cup \{|l(\hat{\theta}|\eta_0) - l(\theta_0|\eta_0)| \geq \epsilon/2\}$, we have

(A.1)
$$P\{|l(\hat{\theta}|\hat{\eta}) - l(\theta_0|\eta_0)| \geq \epsilon\} \leq P\{|l(\hat{\theta}|\hat{\eta}) - l(\hat{\theta}|\eta_0)| \geq \epsilon/2\} + P\{|l(\hat{\theta}|\eta_0) - l(\theta_0|\eta_0)| \geq \epsilon/2\}.$$

On the first term of the right side of (A.1), based on Lemma A.1, Markov's inequality and continuous mapping theorem, one can easily show that given the current value of $\hat{\theta}$, $l(\hat{\theta}|\hat{\eta})$ as a continuous function of η, converges to $l(\hat{\theta}|\eta_0)$ in probability as n goes to infinity. For the second term, under the regularity conditions (C1)–(C4), the log-likelihood function $l(\theta|\eta_0)$ is concave as a function of θ at the true value η_0. Thus the maximum $\hat{\theta}$ exists and is unique and consistent. Based on these two facts, $P\{|\ l(\hat{\theta}|\hat{\eta}) - l(\theta_0|\eta_0)| \geq \epsilon\} \to 0$ as $n \to \infty$, which implies that $\hat{\beta}$ is consistent.

Proof of the asymptotic normality of $\hat{\theta}$. According to the Taylor expansion of $\dot{l}_\theta(\hat{\theta}|\hat{\eta})$ at the true value θ_0, we have

(A.2)
$$\sqrt{n}(\hat{\theta} - \theta_0) = -\{n^{-1}\ \ddot{l}_\theta(\theta_0|\hat{\eta})\ \}^{-1}\ n^{-1/2}\ \dot{l}_\theta(\theta_0|\hat{\eta}) + o_p(1),$$

where $\ddot{l}_\theta(\theta_0|\hat{\eta})$ is the second-order partial derivative of $l(\theta|\hat{\eta})$ at $\theta = \theta_0$, and $-E\{\ddot{l}_\theta(\theta_0|\hat{\eta})\}$ is a positive definite matrix. Thus, to prove the asymptotic normality of $\hat{\beta}$, it is sufficient to prove that the working score function $\dot{l}_\theta(\theta_0|\hat{\eta})$ can be written as the summation of n independent and identically distributed mean zero random variables plus some negligible errors. For this, note that we can rewrite $\dot{l}_\theta(\theta_0|\hat{\eta})$ as

(A.3)
$$\dot{l}_\theta(\theta_0|\hat{\eta}) = I + II\,,$$

where $I = \dot{l}_\theta(\theta_0|\eta_0)$ and $II = \dot{l}_\theta(\theta_0|\hat{\eta}) - \dot{l}_\theta(\theta_0|\eta_0)$.

By following [3, 4], one can easily show that the first term I can be written as the summation of n independent and identically distributed mean zero random variables plus some negligible errors. Next, we prove that the second term II can also be written as the summation of n independent and identically distributed mean zero random variables plus some negligible errors.

Let $\dot{l}_\theta^{(i)}(\theta|\gamma, \Lambda_C(\tilde{C}_i)$ denote the first-order partial derivative of $l^{(i)}(\theta|\gamma, \Lambda_C(\tilde{C}_i))$ about θ. Define

$$\ddot{l}_{\theta\gamma}^{(i)}(\theta|\gamma, \Lambda_C(\tilde{C}_i)) = \frac{\partial}{\partial\gamma}\dot{l}_\theta^{(i)}(\theta|\gamma, \Lambda_C(\tilde{C}_i))\,,$$

and

$$\ddot{l}_{\theta\Lambda_C}^{(i)}(\theta|\gamma, \Lambda_C(\tilde{C}_i)) = \frac{\partial}{\partial s}\dot{l}_\theta^{(i)}(\theta|\gamma, s))|_{s=\Lambda_C(\tilde{C}_i)}\,.$$

Then by using the multivariate Taylor expansion, we can obtain that

$$\begin{aligned} II &= \sum_{i=1}^n \left[\dot{l}_\theta^{(i)}(\theta_0|\hat{\gamma}, \hat{\Lambda}_C(\tilde{C}_i)) - \dot{l}_\theta^{(i)}(\theta_0|\gamma_0, \Lambda_{C0}(\tilde{C}_i))\right] \\ &= \sum_{i=1}^n \Big[\ddot{l}_{\theta\gamma}^{(i)}(\theta_0|\gamma_0, \Lambda_{C0}(\tilde{C}_i))\,(\hat{\gamma} - \gamma_0) \\ &\quad + \ddot{l}_{\theta\Lambda_C}^{(i)}(\theta_0|\gamma_0, \Lambda_{C0}(\tilde{C}_i))\,(\hat{\Lambda}_C(\tilde{C}_i) - \Lambda_{C0}(\tilde{C}_i))\Big] + o_p(1)\,. \end{aligned}$$

To further investigate the equality above, based on [12] and Theorems 8.1–8.3 of [1], we have

$$\hat{\gamma} - \gamma_0 = \frac{1}{n}\sum_{j=1}^n g_j + o_p(n^{-1/2}),$$

where

$$g_j = \Sigma_{\gamma_0}^{-1} \int_0^\tau (Z_j - e(\gamma_0, t)) dM_j(t),$$

$$e(\gamma_0, t) = E\{I(t \le \zeta_j) Z_j e^{\gamma_0^T Z_j}\} / E\{I(t \le \zeta_j) e^{\gamma_0^T Z_j}\},$$

$$M_j(t) = I(C_j \le \zeta_j, C_j \le t) - \int_0^t I(u \le \zeta_j) e^{\gamma_0^T Z_j} d\Lambda_{C0}(u)$$

and $\Sigma_{\gamma_0}^{-1}$ is a positive definite matrix.

Also we have

$$\hat{\Lambda}_C(t) - \Lambda_{C0}(t) = \frac{1}{n} \sum_{j=1}^n b_j(t) + o_p(n^{-1/2})$$

for $\inf\{t : \Lambda_C(t) > 0\} < t < \tau$, where

$$b_j(t) = \int_0^t \frac{dM_j(u)}{E\{I(t \le \zeta_j) e^{\gamma_0^T Z_j}\}} - h(t)^T g_j,$$

and

$$h(t) = \int_0^t e(\gamma_0, u) d\Lambda_{C0}(u).$$

These yield that

$$\begin{aligned} II =& \sum_{i=1}^n \Big\{ \ddot{l}_{\theta\gamma}^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))) E(g_j) \\ &+ \ddot{l}_{\theta\Lambda_C}^{(i)}(\theta_0 | \gamma_0, \Lambda_{C0}(\tilde{C}_i))) E(b_j(\tilde{C}_i)) \Big\} + O_p(1) \\ =& \sum_{i=1}^n d_i(\theta_0, \eta_0) + O_p(1)\,. \end{aligned}$$

It thus follows from (A.3) that the working score function can be written as the summation of n independent and identically distributed mean zero random variables plus some negligible errors such as

$$\dot{l}_\theta(\theta_0 | \hat{\eta}) = \sum_{i=1}^n \{\dot{l}_\theta^{(i)}(\theta_0 | \eta_0) + d_i(\theta_0, \eta_0)\} + O_p(1).$$

Therefore, based on (A.2), $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to a mean zero normal random variable with the covariance matrix $\Sigma = \Gamma' \phi \phi' \Gamma$, where $\Gamma = -\{n^{-1} E\{\ddot{l}_\theta(\theta_0 | \hat{\eta})\}\}^{-1}$ and $\phi = E\{\dot{l}_\theta^{(1)}(\theta_0 | \eta_0) + d_1(\theta_0, \eta_0)\}$. This implies that $\hat{\beta}$ has the asymptotic normality and its asymptotic covariance matrix is the leading $p \times p$ submatrix of Σ.

ACKNOWLEDGEMENTS

The authors wish to thank the editor, Dr. Ming-Hui Chen, and two reviewers for their many helpful and thoughtful comments and suggestions, which greatly improved the paper. This work was partly supported by the National Nature Science Foundation of China Grant Nos. 11471135, 11571133, 11731011, 11671168, and the self-determined research funds of CCNU from the college's basic research of MOE (CCNU15ZD011, CCNU16JCZX11).

Received 2 July 2017

REFERENCES

[1] Fleming, T.R., & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley. MR1100924

[2] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer. MR1777022

[3] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24**, 540–568. MR1394975

[4] Huang, J., & Rossini, A.J. (1997). Sieve estimation for the proportional odds failure time regression model with interval censoring. *Journal of the American Statistical Association* **92**, 960–967. MR1482126

[5] Kalbfleisch, J.D., & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd, ed. New York: Wiley. MR1924807

[6] Lu, M., Zhang, Y., & Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **94**, 705–718. MR2410018

[7] Ma, L., Hu, T., & Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* **102**, 731–738. MR3394289

[8] Nelsen, R.B. (2006). *An Introduction to Copulas*. 2nd, ed. New York: Springer. MR2197664

[9] Ramsay, J.O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425–441. MR1700749

[10] Shih J.H., Louis T.A. (1995). Inference on the association parameter copula models for bivariate survival data. *Biometrics* **51**, 1384–1399. MR1381050

[11] Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer. MR2287318

[12] Wang, M.C., Qin, J. & Chiang, C.T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association* **96**, 1057–1065. MR1947253

[13] Wang, L., Sun, J., & Tong, X. (2008). Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime Data Analysis* **14**, 134–153. MR2398968

[14] Zhao, S., Hu, T., Ma, L., Wang, P., Sun, J. (2015). Regression analysis of informative current status data with the additive hazards model. *Lifetime Data Analysis* **21**, 241–258. MR3324229

[15] Zheng, M., & Klein, J.P. (1995). Estimates of marginal survival for dependent competing risk based on an assumed copula. *Biometrika* **82**, 127–138. MR1332844

Qi Cui
School of Mathematics
Jilin University
Changchun 130012
China
E-mail address: qicui15@mails.jlu.edu.cn

Hui Zhao
School of Mathematics and Statistics &
Hubei Key Laboratory of Mathematical Sciences
Central China Normal University
Wuhan 430079
China
E-mail address: hzhao@mail.ccnu.edu.cn

Jianguo Sun
School of Mathematics
Jilin University
Changchun 130012
China
Department of Statistics
University of Missouri
Columbia, Missouri 65211
USA
E-mail address: sunj@missouri.edu