Semiparametric Copula Models for Bivariate Interval-censored Survival Data

Tao Sun[1,2], Wei Chen[2,1] and Ying Ding[1]

[1]Department of Biostatistics, University of Pittsburgh [2]Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC

Abstract

Bivariate failure time data arise from censored observations of two correlated events, such as eye diseases. We are particularly interested in bivariate "interval-censored" data, due to intermittent assessment times with exact event times unobservable. Copula is a class of parametric association structures that enables flexible associations between margins. In this research, we propose a semiparametric copula-based likelihood approach to model bivariate interval-censored survival data. Specifically, we use a two-parameter Archimedean copula model to account for correlations at both tails and employ Bernstein polynomials to approximate marginal distributions. Then we propose a two-step procedure to estimate model parameters and develop a computationally efficient score test. We perform extensive simulations to evaluate type 1 error and power performance in testing the covariate effect. Finally we apply our model to the GWAS data of a longitudinal study of AMD to identify risk SNPs associated with AMD progression.

Bivariate Interval-censored Data

- (T_1, T_2): bivariate true event times
- $S_j(t_j) = P(T_j > t_j)$, $j = 1, 2$: marginal survival functions
- $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$: joint survival function
- $D_i = \{(L_{ij}, R_{ij}, Z_{ij}, \delta_{ij}), j = 1, 2\}$: observed data
  - $L_{ij} < T_{ij} \leq R_{ij}$: bivariate intervals
  - Z_{ij}: covariates
  - $\delta_{ij} = I(R_{ij} < \infty)$
  - $R_{ij} = \infty$, T_{ij} is right-censored
  - $L_{ij} = 0$, T_{ij} is left-cenosored

Table 1: An example of bivariate interval-censored data

ID	Eye	Left	Right
1	L	5.6	6.4
1	R	5.3	∞
2	L	0.7	3.6
2	R	0	0.7

Copula Model

A two-parameter copula model[1]:

$$C_{\alpha,\kappa}(u, v) = [1 + ((u^{-1/\kappa} - 1)^{1/\alpha} + (v^{-1/\kappa} - 1)^{1/\alpha})^{\alpha}]^{-\kappa}$$

- $\alpha \in (0, 1]$ and $\kappa \in (0, \infty)$
- Kendall's $\tau = 1 - \frac{2\alpha\kappa}{2\kappa+1}$
- $\alpha = 1$ => Clayton copula
- $\kappa \to \infty$ => Gumbel copula

Joint survival function under this copula model:

$$S(t_1, t_2) = C_{\alpha,\kappa}(S_1(t_1), S_2(t_2)), t_1, t_2 \geq 0.$$

The joint likelihood function $L_n((S_1, S_2, \alpha, \kappa); D)$

$$\begin{aligned} &= \prod_{i=1}^{n} \Big[C_{\alpha,\kappa}(S_1(L_{i1}), S_2(L_{i2})) - C_{\alpha,\kappa}(S_1(L_{i1}), S_2(R_{i2})) \\ &- C_{\alpha,\kappa}(S_1(R_{i1}), S_2(L_{i2})) + C_{\alpha,\kappa}(S_1(R_{i1}), S_2(R_{i2})) \Big]^{\delta_{1i}\delta_{2i}} \\ &\times \Big[C_{\alpha,\kappa}(S_1(L_{i1}), S_2(L_{i2})) - C_{\alpha,\kappa}(S_1(R_{i1}), S_2(L_{i2})) \Big]^{\delta_{1i}(1-\delta_{2i})} \\ &\times \Big[C_{\alpha,\kappa}(S_1(L_{i1}), S_2(L_{i2})) - C_{\alpha,\kappa}(S_1(L_{i1}), S_2(R_{i2})) \Big]^{(1-\delta_{1i})\delta_{2i}} \\ &\times C_{\alpha,\kappa}(S_1(L_{i1}), S_2(L_{i2}))^{(1-\delta_{1i})(1-\delta_{2i})} \end{aligned} \quad (1)$$

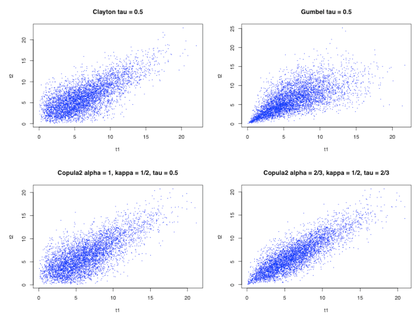

Figure 1: Bivariate survival times under copula models

Esimation and Inference

We implemented the linear transformation model:

$$S_j(t|Z) = \exp(-G\{\exp(Z^T\beta)\Lambda_{0j}(t)\}), j = 1, 2.$$

then approximated $\Lambda_{0j}(t)$ by Bernstein polynomials [2] and got a sieve-based likelihood function. We further proposed a novel two-step estimation procedure similar to [3] to obtain the sieve MLE: $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\kappa}_n, \hat{\beta}_n, \hat{\Lambda}_{n01}, \hat{\Lambda}_{n02})$. We were also able to perform Wald or Score test on covariate effects. We have run extensive simulations to justify the asymptotic properties of our estimates and test statistics.

Simulations

We simulated data from a Clayton Loglogistic model and tested covariate effect under two scenarios:

- Misspecification on margins: Loglog-R, Clayton-W, Clayton-S
- Misspecification on copula: Clayton-L (true), Gumbel-L, Copula2-S

Table 2: Type-I errors

		MSP on Margins					
		Kendall's $\tau = 0.2$			Kendall's $\tau = 0.6$		
MAF	Tail	Loglog-R	Clayton-W	Clayton-S	Loglog-R	Clayton-W	Clayton-S
40%	0.05	0.051	0.103	0.050	0.051	0.131	0.050
	0.01	0.010	0.029	0.010	0.010	0.041	0.010
	0.0001	0.0002	0.0006	0.0002	0.0002	0.0012	0.0001
		MSP on Copula					
		Kendall's $\tau = 0.2$			Kendall's $\tau = 0.6$		
MAF	Tail	Clayton-L	Gumbel-L	Copula2-S	Clayton-L	Gumbel-L	Copula2-S
40%	0.05	0.050	0.061	0.051	0.050	0.065	0.052
	0.01	0.010	0.013	0.010	0.010	0.015	0.010
	0.0001	0.0002	0.0003	0.0002	0.0001	0.0003	0.0001

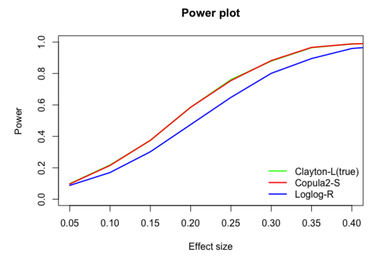

Figure 2: Powers of three tests that control type-I errors well

From Table 2, we found type-I errors were inflated when either copula or margins are misspecified, whereas our models could control the type-I errors well. Our model also achieves comparable powers as the true model. The robust model is less powerful than our model.

GWAS analysis

Multiple large-scale genetic studies had remarkable successes in identifying disease-susceptibility SNPs for AMD [4]. However, the genetic causes for progression to late-AMD, the main cause of blindness, have not been well studied yet. We obtained bivariate interval-censored time-to-late-AMD data for 630 subjects, fit our model and performed score tests on SNPs. The top 5 SNPs associated with late-AMD progression are shown in table 3.

Table 3: Top 5 SNPs in ARMS2 region on chromosome 10

SNP	MAF	p-value
rs2293870	0.29	1.05E-06
rs1049331	0.29	1.06E-06
rs2284665	0.29	1.12E-06
rs11200638	0.28	1.16E-06
rs58077526	0.31	1.23E-06

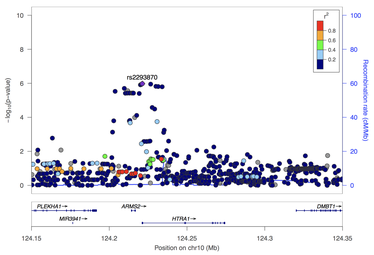

Figure 3: SNP positions

Conclusions

In summary, we have proposed a two-parameter copula-based semi-parametric likelihood approach for regression analysis of bivariate interval-censored survival data. Simulations suggestion that type-I errors and powers of our score test are well controlled, even comparable to the true model. We have built an R package 'CopulaTest' that performs estimations and tests (Wald and Score) under various copula models (Clayton, Gumbel, two-parameter copula) and marginal assumptions (Weibull, Loglogistic, Sieve). The package also outputs AIC for model selection purpose. Finally, using our novel copula model, we were able to identified significant SNPs associated with progression to late-AMD in a large-scale GWAS study.

References

[1] Z. Chen, Ph.D thesis (2012).

[2] Q Zhou, T Hu, J Sun, JASA **112** (2017), 664-672.

[3] L Sun, L Wang, J Sun, Scand. J. Stat **33**, (2006).

[4] L Fritsche et.al, Nat. Genet, **2** (2016), 134-143.

Acknowledgements

This work is supported by the research grant R01EY024226 (PI: Chen W.) from NEI.