DEPENDENT CENSORING IN SURVIVAL REGRESSION MODEL

Radoslav Kovář – Ivana Malá – Filip Habarta

Abstract

Survival models are often based on the assumption of independence between survival time and censoring time. This paper explores the performance of certain parametric survival models in cases where this assumption doesn't hold. Firstly, a correctly specified likelihood for a bivariate parametric copula and two parametric marginal survival functions is derived. Secondly, the bias and variance of this estimator are compared on simulated data with the bias and variance of the standard estimator. The results of this simulation are confirmed analytically for special cases. In the third part, we discuss the consequences for regression models using a dummy covariate. In particular, we present a model where the parameters of dependence are also a function of covariate values. We show that standard survival models can lead to flawed conclusions if independence between survival time and censoring time is assumed wrongly, even in cases when parametric families of the margins are correct.

Key words: survival analysis, dependent censoring, copula functions

JEL Code: C41, C15

Introduction

Analysis of time-to-event data has a special place among regression models. One of its distinctive features is the presence of censored observations. If a study ends before the event of interest occurs for a given observation, one should still make use of precious information – namely that its time of interest is greater than the study time. Under the assumption of independence between these two times, such a right-censored observation can easily be included in the likelihood and an estimator with desirable properties can be derived.

However, this assumption of independence isn't tenable for more complicated censoring mechanisms. Kalbfleisch and Prentice (2003, p. 249) analyze the competing dependent risks of leukemia relapse and graft versus host disease. One can also question the independence of time until default and time until early repayment of a given loan (Stepanova

and Thomas, 2002). It is therefore useful to simulate the behavior of standard estimators in these settings and compare them with more appropriate ones.

In the first section of this paper, we derive a likelihood function which accounts for dependent censoring, we specify the scope of this study and describe the simulation setting. We begin the second section by comparing the bias, variance and mean squared error (MSE) of our estimator with the standard one on simulated data. Then, we discuss the consequences of these simulations for regression models.

This paper builds especially on studies by Emura and Chen (2016) and Li et al. (2007).

1 Methodology

1.1 Notation

Denote the time of interest as D, the censoring time as E, their minimum as T and the event $I(D < E)$ as δ. Let f_X, S_X, h_X be the density, survivor function and hazard function of a random variable X depending on a parameter θ_X where $S_X(x) = \int_x^\infty f(u)du$ and $h_X(x) = \frac{f_X(x)}{S_X(x)}$. Finally, $C(c_1, c_2)$ stands for values of a copula function defined on the square $[0;1]\times[0;1]$, which depend on the parameter τ and increase for each argument from 0 to 1. Other properties of copulas can be found in Nelsen (2006).

1.2 Likelihood under dependent censoring

If the joint distribution of D and E is derivable, then there exists according to Sklar's theorem (1959) a derivable copula such that:

$$P(D > d, E > e) = C\left(S_D(d), S_E(e)\right). \tag{1}$$

The distribution is therefore completely determined by its margins and the copula structure. Then, the conditional density equals:

$$f_{E|D}(e|d) = \frac{f_{D,E}(d,e)}{f_D(d)} = \frac{\frac{P(D > d, E > e)}{\partial d \partial e}}{f_D(d)} = f_E(e)\frac{\partial^2 C\left(S_D(d), S_E(e)\right)}{\partial c_1 \partial c_2} \tag{2}$$

and the conditional survivor function equals:

$$S_{E|D}(e|d) = \int_e^\infty f_{E|D}(u|d)du = \frac{\partial C\left(S_D(d), S_E(e)\right)}{\partial c_1}. \tag{3}$$

If the event of interest is observed for the unit i at the time t_i then $\delta_i = 1$ and its contribution to the likelihood equals:

$$\mathrm{L}_i = \int_{t_i}^{\infty} f_{D,E}(t_i, u)du = f_D(t_i)S_{E|D}(t_i|t_i). \tag{4}$$

Since the contribution of censored observations is symmetrical, the log-likelihood of n observations equals:

$$ln\mathrm{L} = \sum_{i=1}^{n} \delta_i (lnf_D(t_i) + ln\frac{\partial C\left(S_D(t_i), S_E(t_i)\right)}{\partial c_1}) + (1-\delta_i)(lnf_E(t_i) + ln\frac{\partial C\left(S_D(t_i), S_E(t_i)\right)}{\partial c_2}). \tag{5}$$

We denote the estimators which maximize $ln\mathrm{L}$ as $\hat{\theta}_D^{DEP}, \hat{\theta}_E^{DEP}, \hat{\tau}$. For the special case of the independence copula $C(c_1, c_2) = c_1 c_2$, the equation (5) can be split in two parts:

$$ln\mathrm{L} = \sum_{i=1}^{n} \delta_i lnf_D(t_i) + (1-\delta_i)lnS_D(t_i) + \sum_{i=1}^{n}(1-\delta_i)lnf_E(t_i) + \delta_i lnS_E(t_i) = ln\mathrm{L}_D + ln\mathrm{L}_E. \tag{6}$$

Since one is usually interested only in the estimation of θ_D and $ln\mathrm{L}_E$ doesn't contain any information about it, it suffices to maximize $ln\mathrm{L}_D$. We denote the commonly used estimators which maximize $ln\mathrm{L}_D$ and $ln\mathrm{L}_E$ as $\hat{\theta}_D^{INDEP}, \hat{\theta}_E^{INDEP}$.

For different copulas and marginal distributions, we always performed a numerical maximization of $ln\mathrm{L}$, using $\hat{\theta}_D^{INDEP}, \hat{\theta}_E^{INDEP}$ and $\tau = 0$ as starting points. We applied the algorithms by Nelder and Mead (1965) and by Byrd et al. (1996) and haven't detected any irregular behavior of $ln\mathrm{L}$ in the neighborhood of the starting points. All simulations build on `R` packages `survival` by Therneau (2015) and `copula` by Hofert et al. (2017).

1.3 Purpose of the simulations

If one doesn't have any prior idea about $h_D(t)$ and the censoring distribution, a common empirical estimate $\hat{h}_D^*(t)$ is based on the proportion of observed events just after t among observations still at risk at t:

$$h_D^*(t) = \lim_{dt\to 0}\frac{1}{dt}P(t < D \le t + dt, D < E \,|\, D > t, E > t) = \frac{\lim_{dt\to 0}\frac{1}{dt}\int_t^{t+dt} f_D(u)S_{E|D}(u|u)du}{P(D > t, E > t)}$$
$$= \frac{f_D(t)}{S_D(t)} \times \frac{S_D(t)}{C\left(S_D(t), S_E(t)\right)}\frac{\partial C\left(S_D(t), S_E(t)\right)}{\partial c_1} := h_D(t) \times r(t). \tag{7}$$

For the special case of the independence copula $C(c_1, c_2) = c_1 c_2$, one really obtains $h_D^*(t) = h_D(t)$. However, for dependent D and E the estimator is biased by the factor $r(t)$. Emura and Chen (2016) have already extensively studied the behavior of $\hat{h}_D^*(t)$ for different parametric settings.

Our focus is different. We assume one already has a correct idea about the parametric families of D and E (based on previous research, the nature of the data-generating process, etc.). We further assume that one believes for the same reasons in the independence of D and E and uses therefore the standard estimator $\hat{\theta}_D^{INDEP}$. This study shows that being mistaken about the independence might lead to bias even if the families of the margins are correct. Our simulations are therefore relevant especially in situations where parameters of the copula change over time, see Barthel et al. (2018).

1.4 Setting of the simulations

Archimedean copulas of the form

$$C(c_1, c_2) = \psi(\psi^{-1}(c_1) + \psi^{-1}(c_2)) \tag{8}$$

are considered classical by Nelsen (2006) and were also used in the simulation study by Emura and Chen (2016) mentioned above. Their derivatives are $\frac{\partial C(c_1, c_2)}{\partial c_k} = \frac{(\psi^{-1})'(c_k)}{(\psi^{-1})'(C(c_1, c_2))}$. Two specific families were chosen for the simulations: The Frank (1978) copula with the generating function:

$$\psi(u) = -\frac{\ln(1-(1-e^{-\tau})e^{-u})}{\tau} \tag{9}$$

and the Ali-Mikhail-Haq (AMH) copula (Ali et al., 1978) with the generating function:

$$\psi(u) = \frac{1-\tau}{e^u - \tau}. \tag{10}$$

The parameter τ is related to Kendall's tau and both copulas permit to model both positive and negative correlation. While Frank copulas are symmetrical, one tail of AMH copulas is heavier.

Two distributions were chosen for D and E: the exponential distribution with the survivor function parameterized as $S_D(d) = \exp(-d\exp(-\theta_D))$ and the log-logistic distribution with scale 1 parameterized as $S_D(d) = (1 + d\exp(-\theta_D))^{-1}$. Both distributions are commonly used in parametric survival analysis, see Hosmer and Lemeshow (2008).

Moreover, there exists a closed-form solution for $\hat{\theta}_D^{INDEP}$ in the case of exponential distribution, which allows us to confirm its bias analytically:

$$\hat{\theta}_D^{INDEP} = -\ln \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} \xrightarrow[n \to \infty]{} -\ln \frac{P(D < E)}{E(T)}$$

$$= -\ln \frac{\int_0^{\infty} f_D(u) S_{E|D}(u|u) du}{\int_0^{\infty} u f_D(u) S_{E|D}(u|u) du + \int_0^{\infty} u f_E(u) S_{D|E}(u|u) du} := b_D(\theta_D, \theta_E, \tau) \tag{11}$$

Since $\hat{\boldsymbol{\theta}}^{DEP} = \left[\hat{\theta}_D^{DEP}, \hat{\theta}_E^{DEP}, \hat{\tau}\right]$ is consistent, one can make for large enough n the approximation $\hat{\theta}_D^{INDEP} \approx b_D(\hat{\theta}_D^{DEP}, \hat{\theta}_E^{DEP}, \hat{\tau})$. Delta method then yields an approximate relationship between the covariance matrix of $\hat{\boldsymbol{\theta}}^{INDEP} = \left[\hat{\theta}_D^{INDEP}, \hat{\theta}_E^{INDEP}\right]$, which has a closed-form solution, and the covariance matrix of $\hat{\boldsymbol{\theta}}^{DEP}$:

$$Var\left[\hat{\boldsymbol{\theta}}^{INDEP}\right] \approx Var\left[\mathbf{b}(\boldsymbol{\theta}) + \mathbf{B}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}^{DEP} - \boldsymbol{\theta}\right)\right] = \mathbf{B}(\boldsymbol{\theta}) Var\left[\hat{\boldsymbol{\theta}}^{DEP}\right] \mathbf{B}(\boldsymbol{\theta})^T, \tag{12}$$

where $\mathbf{B}(\boldsymbol{\theta})$ is the Jacobian of b_D and b_E evaluated at the true parameters. In this case, one can therefore also confirm the variance and thus the MSE analytically.

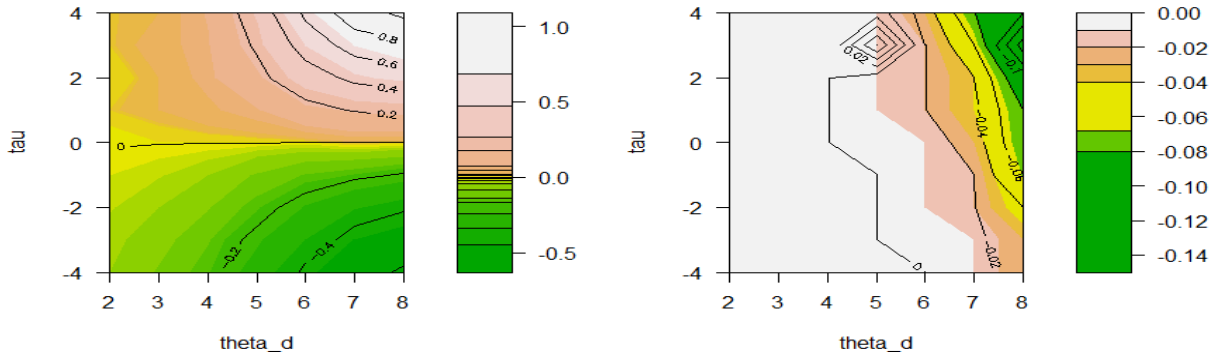
Four simulation settings based on two copulas and two marginal distributions described above are presented in the next section. Within each setting, we fixed the parameter $\theta_E = 5$, varied the parameter θ_D from 2 to 8 and varied the parameter τ from -4 to 4 for the Frank copula and from -0.75 to 0.75 for the AMH copula. The sign of τ corresponds with the sign of the correlation between D and E and with increasing θ_D one progresses from mild to heavy censoring. These values were chosen to assure that one always observes a mixture of censored and uncensored observations. For each combination of parameters we generated a sample of 1000 observations and calculated $\hat{\theta}_D^{INDEP}$ and $\hat{\theta}_D^{DEP}$. This was repeated 100 times to learn about the distributions of both estimators and therefore also about their bias, variance and MSE.

2 Results of the simulations

2.1 Simulations without covariates

The following figures refer to the setting with Frank copula and exponential margins. The first figure shows the bias of $\hat{\theta}_D^{INDEP}$, the second compares the variance of $\hat{\theta}_D^{INDEP}$ and $\hat{\theta}_D^{DEP}$:

Fig. 1: Bias of the standard estimator, comparison of variances of both estimators



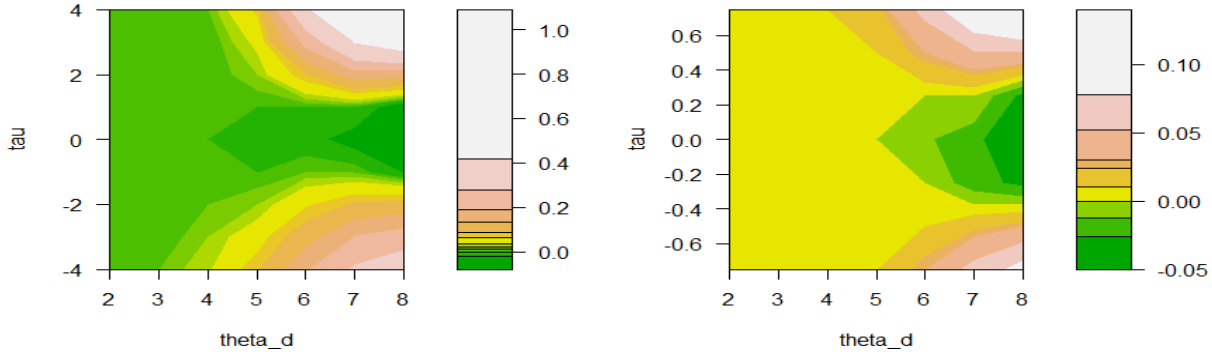
Source: the authors.

Black solid lines show the values predicted by (11) and (12).

The first figure shows that the estimator $\hat{\theta}_D^{INDEP}$ is too optimistic for positively correlated D and E and vice versa. The absolute value of the bias increases with the amount of censoring, but up to 50% censoring ($\theta_D = 5$), the size of the bias is negligible. In contrast, $\hat{\theta}_D^{DEP}$ maximizes the correctly specified likelihood and is therefore always asymptotically unbiased. However, the second figure indicates that its variance is higher, because (5) and (6) have the same information from the sample and (5) always spends part of it to estimate an additional parameter τ. Predicted values match well with the simulations.

These results can be used to compare the MSE of both estimators. The left figure below refers to the same setting as previously, the right to the setting with AMH copula and log-logistic margins:

Fig. 2: Comparison of MSE of both estimators for different settings



Source: the authors.

Identical results were obtained for the remaining two settings.

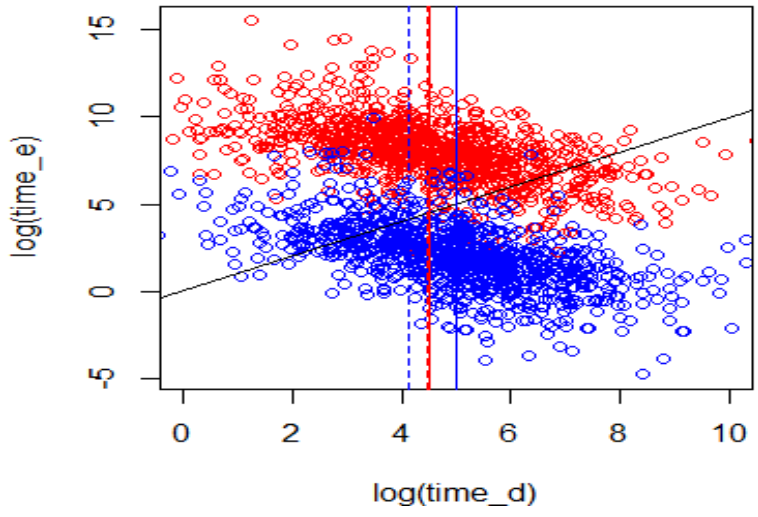
For mild censoring, our estimator is slightly preferable to $\hat{\theta}_D^{INDEP}$. For heavy censoring with strong dependence, we have $MSE(\hat{\theta}_D^{DEP})$ $MSE(\hat{\theta}_D^{INDEP})$. Interestingly, the standard estimator performs slightly better than ours for heavily censored data with very weak dependence, despite being incorrectly specified. In these situations, the reduction in variance overcompensates for the bias of $\hat{\theta}_D^{INDEP}$. The exact size of these differences depends on the simulation setting.

2.2 Regression models

The simulations can be adjusted to derive estimated effects in a regression model with a dummy covariate **x**.

The following figure is based on the setting with Frank copula ($\tau = -4$) and log-logistic margins. Red points ($x_i = 0$) were generated from a distribution with the margins $\theta_D^{x=0} = 4.5$ and $\theta_E^{x=0} = 8$, blue points ($x_i = 1$) were generated from a distribution with the margins $\theta_D^{x=1} = 5$ and $\theta_E^{x=1} = 2$. The values θ_D are denoted with solid vertical lines. Points under the diagonal are censored.

Fig. 3: Bias in regression models



Source: the authors.

The dashed lines correspond to the estimates $\hat{\theta}_D^{x=0,INDEP}$ and $\hat{\theta}_D^{x=1,INDEP}$. Because of the negative correlation, both underestimate the true values. However, the size of the bias is bigger for $x = 1$ due to heavier censoring. In this case, it even makes the positive effect $\theta_D^{x=1} - \theta_D^{x=0} = 0.5$

appear negative and "statistically significant". Similar results can be derived for positively correlated data and other simulation settings.

On the contrary, our estimators are unbiased even in cases when τ is a function of **x**. The following table shows the estimates in the same setting as above except for $\tau^{x=0} = 4$:

Tab. 1: Dependence as a function of covariates

Parameter	$\theta_D^{x=0}$	$\theta_E^{x=0}$	$\tau^{x=0}$	$\theta_D^{x=1}$	$\theta_E^{x=1}$	$\tau^{x=1}$
True value	4.5	8	4	5	2	-4
Estimate	4.5044	8.3963	2.6820	4.9493	1.9986	-3.8157
(st.error)	(0.0548)	(0.4441)	(1.1505)	(0.1396)	(0.0570)	(0.7400)

Source: the authors.

These results can be easily generalized for multiple and continuous covariates.

Conclusion

This simulation study illustrates the importance of accounting for dependent censoring even in cases when one knows the parametric families of the marginal distributions. Ignoring this issue might lead to flawed conclusions especially for heavily censored data. This confirms and extends the results by Emura and Chen (2016) and Li et al. (2007). We derived a new estimator based on the correctly specified likelihood and show its satisfactory performance for Archimedean copulas.

Our estimator is based on the strong assumption that one has a correct idea about the parametric family of the copula. In further research we relax this assumption and find the optimal copula under several possible ones. Moreover, we extend our findings to a mixture cure survival model and apply the methodology in the field of credit scoring.

Acknowledgment

This paper is supported by the grant F4/80/2018 (Společné rozdělení dvou dob přežití a jeho změny v čase) which has been provided by the Interní grantová agentura Vysoké školy ekonomické v Praze (Internal Grant Agency of the University of Economics in Prague).

References

Ali, M. M., Mikhail, N., & Haq, M. (1978). A class of bivariate distributions including the bivariate logistic.*Journal of Multivariate Analysis,8*(3), 405-412. doi:10.1016/0047-259x(78)90063-5

Barthel, N., Geerdens, C., Killiches, M., Janssen, P., & Czado, C. (2018). Vine copula based likelihood estimation of dependence patterns in multivariate event time data.*Computational Statistics & Data Analysis,117*, 109-127. doi:10.1016/j.csda.2017.07.010

Byrd, R., Peihuang, L., & Nocedal, J. (1996). A limited-memory algorithm for bound-constrained optimization. doi:10.2172/204262

Emura, T., & Chen, Y. (2016). Gene selection for survival data under dependent censoring: A copula-based approach.*Statistical Methods in Medical Research,25*(6), 2840-2857. doi:10.1177/0962280214533378

Frank, M. J. (1978). On the simultaneous associativity ofF(x,y) andx y -F(x,y).*Aequationes Mathematicae,18*(1-2), 266-267. doi:10.1007/bf01844082

Marius Hofert, Ivan Kojadinovic, Martin Maechler and Jun Yan (2017). copula: Multivariate Dependence with Copulas. R package version 0.999-18 URL https://CRAN.R-project.org/package=copula

Hosmer, D. W., Lemeshow, S., & May, S. (2008).*Applied survival analysis: Regression modeling of time to event data*. New York: Wiley.

Kalbfleisch, J. D., & Prentice, R. L. (2003).*The statistical analysis of failure time data*. New York: J. Wiley.

Li, Y., Tiwari, R. C., & Guha, S. (2007). Mixture cure survival models with dependent censoring.*Journal of the Royal Statistical Society: Series B (Statistical Methodology),69*(3), 285-306. doi:10.1111/j.1467-9868.2007.00589.x

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization.*The Computer Journal,7*(4), 308-313. doi:10.1093/comjnl/7.4.308

Nelsen, R. B. (2006).*An introduction to copulas*. New York: Springer.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8: 229–231

Stepanova, M., & Thomas, L. (2002). Survival Analysis Methods for Personal Loan Data.*Operations Research*,*50*(2), 277-289. doi:10.1287/opre.50.2.277.426

Therneau T (2015). *A Package for Survival Analysis in S*. version 2.38, https://CRAN.R-project.org/package=survival.

Contact

Radoslav Kovář

University of Economics, Winston Churchill Sq. 1938/4, Prague, Czech Republic

xkovr11@vse.cz

Ivana Malá

University of Economics, Winston Churchill Sq. 1938/4, Prague, Czech Republic

malai@vse.cz

Filip Habarta

University of Economics, Winston Churchill Sq. 1938/4, Prague, Czech Republic

xhabf00@vse.cz