'Tests of functions

Jeppe Ekstrand Halkjær Madsen

2019-04-16

Testing for independence

Estimation of Kendall's τ for censored data

The function `tauCens` can estimate Kendall's τ for censored data. This relies on some extrapolation which here is done in a very simple way using the Kaplan-Meier estimator. This means that the extrapolation doesn't use information from the other failure time when extrapolating, which makes the estimator biased towards 0. The standard error estimate uses an assumption of independence and only takes the variance of the nominator into account, which means that the estimate of the standard error is going to be too small on average. These problems are illustrated with the following simulation. 100 pairs of failure times are simulated from the Clayton copula and the Gumbel copula respectively. The marginal distributions are exponential with a scale parameter of 1, while censoring times are exponentially distributed with a scale parameter of 2 leading to roughly 30 % censoring. Then Kendall's $\hat{\tau}$ and its standard error are estimated. The simulation is repeated 10000 times and is done for four different values of Kendall's τ: 0, 0.25, 0.5, 0.75. The results are summarised in Table 1.

τ	0	0.25	0.5	0.75
Clayton	0.007	0.173	0.322	0.440
Clayton theoretical SE	0.035	0.035	0.035	0.035
Clayton empirical SE	0.054	0.070	0.095	0.120
Gumbel	0.007	0.134	0.272	0.407
Gumbel theoretical SE	0.035	0.035	0.035	0.035
Gumbel empirical SE	0.054	0.065	0.085	0.107

Table 1: Average estimate of Kendall's τ for different copulas and true values of τ

The estimator looks unbiased in the case of $\tau = 0$ (which corresponds to the independence copula so the results should be the same for both copulas in this case, which they are), but otherwise biased as expected. The bias seems to be worse for the Gumbel copula than for the Clayton copula. One theory could be that this is because the Gumbel copula has upper tail dependence while the Clayton copula has lower tail dependence so the dependence gets "censored away" in the case of the Gumbel copula, but not in the case of the Clayton copula. We also see that the empirical standard error on average is greater than the estimated one. The estimated standard error is unaffected by the value of Kendall's τ while the actual standard error is bigger the more Kendall's τ differs from 0.

Conclusion

The estimator is far from perfect, but can still be useful as a quick way of getting an impression of how much (if any) dependence exists in the data.

Estimating tail dependence for censored data

Lower and upper tail dependence between two random variables, X_1 and X_2, are defined as

$$\lambda_l = \lim_{q \downarrow 0} P(X_1 \leq F_1^{\leftarrow}(q) | X_2 \leq F_2^{\leftarrow}(q)), \qquad \lambda_u = \lim_{q \uparrow 1} P(X_1 > F_1^{\leftarrow}(q) | X_2 > F_2^{\leftarrow}(q)),$$

where F_1 and F_2 are the marginal distribution functions and $f^{\leftarrow}(x)$ is the generalized inverse of f. It is difficult to estimate the limits so what is done in this package is that the probabilities

$$P(X_1 \leq F_1^{\leftarrow}(q)|X_2 \leq F_2^{\leftarrow}(q)), \qquad P(X_1 > F_1^{\leftarrow}(q)|X_2 > F_2^{\leftarrow}(q)),$$

are estimated, for some value of q "close" to 0 or 1. These probabilities are estimated in two different ways in this package:

One way of estimating the probabilities is by rewriting them in terms of survival functions:

$$\lambda_l = \frac{1 - P(X_1 > F_1^{\leftarrow}(q)) - P(X_2 > F_2^{\leftarrow}(q)) + P(X_1 > F_1^{\leftarrow}(q), P(X_2 > F_2^{\leftarrow}(q)))}{q},$$
$$\lambda_u = \frac{P(X_1 > F_1^{\leftarrow}(q), P(X_2 > F_2^{\leftarrow}(q)))}{1-q}.$$

These probabilities can be estimated by using the Kaplan-Meier estimator for the univariate survival functions and the Dabrowska estimator for the bivariate survival function. This is a very legitimite way of doing it, but can be slow since the Dabrowska estimator can take a lot of time for big datasets (number of entries to calculate is n^2, where n is the number of bivariate failure times).

There is, however, a faster and simpler way to estimate the probabilities: Imagine we have a dataset of bivariate failure times and event indicators $(t_{i1}, t_{i2}, \delta_{i1}, \delta_{i2})_{i=1,\ldots,n}$. Take the subset of (t_{i1}) where $t_{i2} \leq \hat{F}_2^{\leftarrow}$ and $\delta_{i2} = 1$, where $\hat{F}$ is 1 minus the Kaplan-Meier estimate of the survival function ($\delta_{i2} = 1$ is not required for upper tail dependence, and the inequality is of course turning the other way). Then the probabilities are estimated on this subsample using the Kaplan-Meier estimator.

How do they work? Both estimators have been tested for both upper and lower tail dependence on simulated data. $q = 0.10$ for lower tail dependence and $q = 0.80$ for upper tail dependence. 100 bivariate survival times were simulated from two copula models: the Clayton copula and the Gumbel copula. Marginal distributions were exponential with a parameter of 1, while censoring times were simulated from an exponential distribution with a rate parameter of 0.3 leading to roughly 23 % censoring. The results are summarized in Table 2 with the average estimated tail dependence from 10000 simulations.

Copula (tail)	Method	$\tau = 0$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
Clayton (lower)	True value	0.100	0.419	0.709	0.891
	Dabrowska	0.103	0.421	0.708	0.890
	Fast	0.147	0.424	0.714	0.894
Gumbel (lower)	True value	0.100	0.203	0.385	0.647
	Dabrowska	0.104	0.213	0.391	0.651
	Fast	0.145	0.232	0.387	0.645
Clayton (upper)	True value	0.200	0.294	0.430	0.648
	Dabrowska	0.185	0.273	0.398	0.607
	Fast	0.147	0.221	0.334	0.535
Gumbel (upper)	True value	0.200	0.435	0.647	0.835
	Dabrowska	0.184	0.411	0.615	0.790
	Fast	0.150	0.346	0.549	0.728

Table 2: Average estimated tail dependence from a few different scenarios.

The estimators generally seem to underestimate upper tail dependence, but work well for lower tail dependence.

The idea with this function is that it makes it possible to get a feeling of whether the data set has lower or upper tail dependence. This can be useful when selecting a specific model. If we estimate a value of Kendall's τ of around 0.25 and then estimate a lower tail dependence of 0.4 at the 10 % quantile then it would be wise not to use the Gumbel copula to describe your data.