MODELING OF BIVARIATE DATA FOR DEPENDABILITY

Dr.ing. Adrian Stere PARIS, Dr. Constantin TÂRCOLEA
Univ. Politehnica Bucharest
email: adrian.paris@upb.ro, constantin_tarcolea@yahoo.com

Abstract. *For technical applications the dependability generally measures availability, reliability, maintainability, and connected activities like costs. The common effect of two or many properties on the same objects can be estimated with multivariate distributions using copula structures.*

Keywords: *bivariate data, rank, Clayton copula*

1.Introduction

In applied reliability it is important to establish the joint influence of the factors of the process under study. This paper investigated the dependence features using bivariate rank and Clayton copula distributions. In the paper "Maintenance Dependence Modeling with Gaussian Copulas" [5] we modeled a bivariate cumulative normal distribution - of two controlled variables, mileage (km) and maintenance costs. For the measure of association between the components of the random variables it was used Kendal's coefficient correlation tau [6].

"Copula – based models are natural situations where learning about thr association between the variables is important, since the effect of the dependence structure is then easily separated from that of the marginals"[3].

The paper modeled a joint bivariate cumulative distribution of two controlled variables, mileage and maintenance costs, once, based on rank model and second with copula functions. The measure of non-linear connection between these two choosing independent variables is estimated by the Kendall's correlation coefficient tau.

2. Rank copula

Nonparametric method uses essentially as mathematical object the empirical CDF as marginal distributions and for the bivariate distribution too. The univariate empirical CDF is given by:

$$\hat{F}_i(x) = \frac{1}{n+1}\sum_{t=1}^{n} 1_{\{x_{t,i}\leq x\}} \qquad (1)$$

where **1** is the indicator function.

At the beginning it is not supposed (assumption) hypotheses for the distributions of margins. Initial it is calculated the univariate empirical marginal values of the observed repartition of data. Let a sample (x=milage,y=costs)=(x_1,y_1; ...;x_i, y_i; ...; x_n,y_n) [4].

The observed data were rescaled, namely: $\hat{u}_i = \frac{Rank(x_i)}{n+1}$, respectively: $\hat{v}_i = \frac{Rank(y_i)}{n+1}$.

The vector ($\hat{\boldsymbol{u}}$,$\hat{\boldsymbol{v}}$)=($\hat{u}_1$,$\hat{v}_1$; ...;$\hat{u}_i$,$\hat{v}_i$; ...; $\hat{u}_n$,$\hat{v}_n$) is an approximate sample for the rank based bivariate empirical cumulative function[1]:

$$C_n(\hat{u},\hat{v})= \frac{1}{n+1}\sum_{i=1}^{n} I(U_i \leq \hat{u}, V_i \leq \hat{v}) \qquad (2)$$

where $(\hat{u},\hat{v}) \in [0;\ 1]x[0;\ 1]$, and I(.) is the indicator function, taking the value 1 if $U_i \leq \hat{u}, V_i \leq \hat{v}$ and 0 otherwise.

The 2-dimensional copula $C_n(\hat{u},\hat{v})$ is a bivariate distribution C with uniform marginals on [0; 1].

A consistent estimation can be described as the distribution function of the sample of the normalized ranks.

To plot data points on three axes [6] to visualize the relationship between three variables are used 3D scatter plots (fig.1). Because the correlation between variables is high the points are close to making a specific form, inducing the construction of a copula. In the 3D scatter plot below, mileages, costs, and there dependability are plotted against each other (fig.2) for the given field data set.

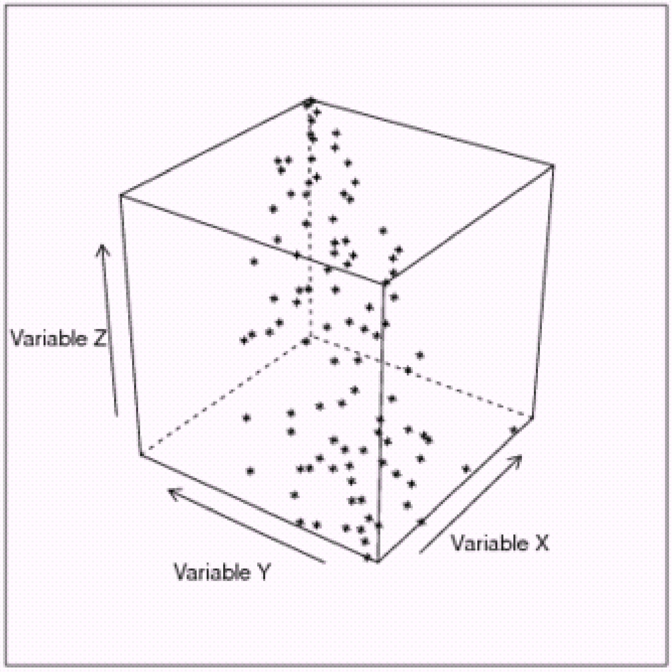

Fig.1 The 3D scatter plot of the values of the rank CDF

The pattern of the data plot suggests a particular cloud of points, compatible with dependence copula models, particularly a Clayton copula. A detailed analysis of pairs dependencies is illustrated in figure 2.

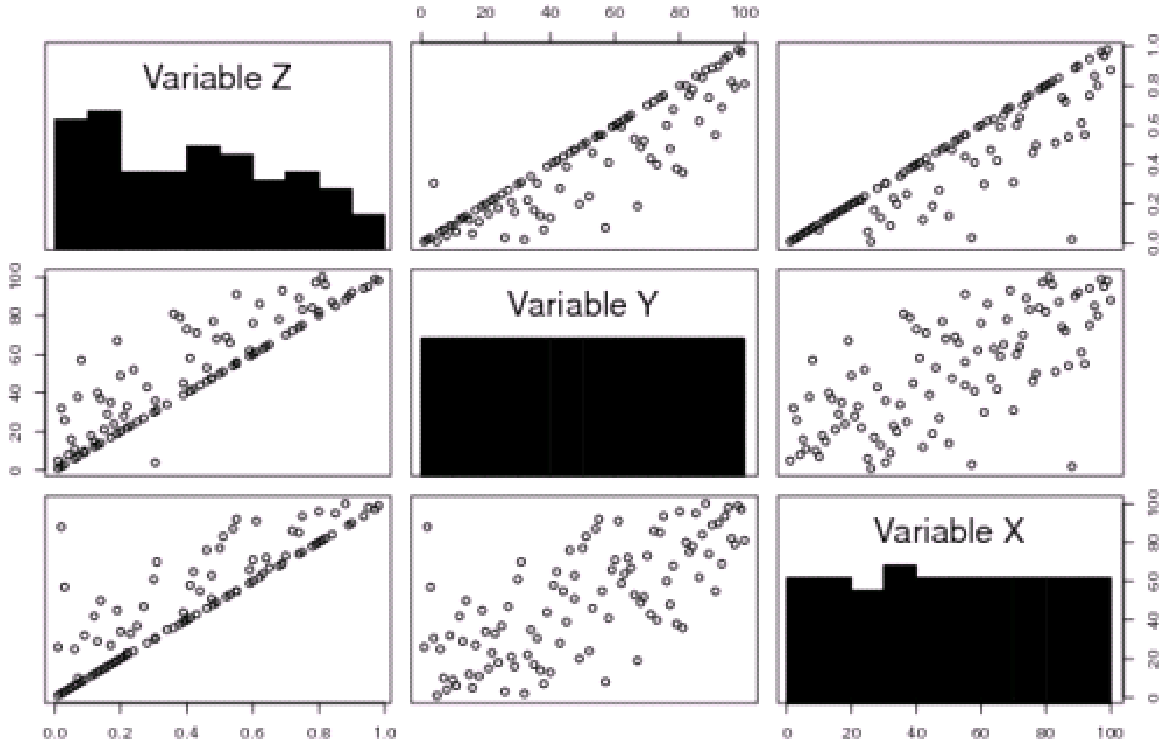

Fig.2 The paired dependence between X (mileage), Y (costs), Z (rank CDF)

3. Clayton Copula

In the first step it was obtained two uniform distributions on [0;1] using the values of the univariate cumulative normal distribution function, (u,v). Next we used the bivariate family of Clayton copula, C(u,v,θ), [2], were:

$$C(u,v,\theta) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}} \tag{3}$$

$\theta = \frac{2\tau}{1-\tau}$, and τ is the Kendal's correlation coefficient between the independent variables miles and costs. Because τ = 0.5037 it results an estimation of $\hat{\theta}$ =2.03226. With the copula:

$$C(u,v,2.03226) = (u^{-2.03226} + v^{-2.03226} - 1)^{-\frac{1}{2.03226}} \tag{4}$$

were calculated the set of 100 values for the bivariate cumulative distribution function.

4. Conclusions

At the beginning it was evaluated the dependence between the independent variables miles and costs with Pearson correlation coefficient, r = 0.6588; Spearman correlation coefficient, rs = ρ(rho) = 0.6964; Kendall correlation coefficient, τ(tau) = 0.5037 (p-value = 0.00 for each coefficient) [7]. These values prove that the association between the variables is statistical significant. Furthermore the values of the bivariate cumulative distribution function have been developed with rang correlation and copula calculus. The correlation between the first component and the empirical rank distribution function indicates a very good link: Pearson coefficient r = 0.9852, Spearman coefficient rs = ρ(rho) = 0.9887 Kendall coefficient τ(tau) = 0.9330. Analog was studied the relation between the second component and the empirical rank distribution function with similar results (Pearson coefficient is r = 0.9022; Spearman coefficient rs = ρ(rho) = 0.8952; Kendall coefficient τ(tau) = 0.7443. The Clayton copula for the bivariate dependence cumulative function, also a useful model gives, in comparison with the rang model, following results: Pearson correlation coefficient r = 0.9852; Spearman correlation coefficient rs = ρ(rho) = 0.9887; Kendall correlation coefficient τ(tau) = 0.9330, what confirmed a closed measure of association. For the reliability of the chose models it was calculated the correlation between the rang cumulative function and CDF Weibull estimate (Pearson coefficient r = 0.8663; Spearman coefficient rs = ρ(rho) = 0.8844; Kendall coefficient τ(tau) = 0.7304). Finally was estimated the dependence between the Clayton copula values and CDF Weibull (Pearson coefficient r = 0.8647; Spearman coefficient rs = ρ(rho) = 0.9084; Kendall coefficient τ(tau) = 0.7834.

References

1. **Amblard C, Girard S.,** *Symmetry and Dependence Properties within a Semiparametric Family of Bivariate Copulas*. Nonparametric Statistics, 14 (6): 2002, pp.715-727.

2. **Clayton, D. G.,** *'A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence'*. Biometrika 65, 1978, pp. 141–151

3. **Genest C., Ghoudi K., Rivest L.-P.**, *A semiparametric estimation procedure of dependence parameters in multivariate families of distributions*, Biometrika, Volume 82, Issue 3, 1 September 1995, pp. 543–552.

4. **Paris A. S., Târcolea C., Andreescu C., Dragomirescu C.**, *Maintenance costs statistics for urban cars*, Proceedings of the European Automotive CongressEAEC-ESFA 2015,Vol. I, pp. 297-305, Springer Verlag, 2015, http://www.eaec-esfa2015.com/index.php/eaec/2015/paper/ view/84

5. **Paris A. S., Sylvan D., C. Târcolea C.,** *Maintenance Dependence Modeling with Gaussian Copulas,* - Conferinta Internationala "Calitate si Siguranta in Functionare" CCF 2016, 14-16 septembrie 2016, Revista ASIGURAREA CALITATII – QUALITY ASSURANCE, Iulie – Septembrie 2016 Anul XXII, Numãrul 87, pp. 28-32

6. **Wessa, P.**, *Free Statistics Software,* Office for Research Development and Education, version 1.2.1, 2018, URL https://www.wessa.net/

7. ***https://www.answerminer.com/calculators/correlation-test