Journal of Applied Statistics

Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Regression in a copula model for bivariate count data

Aristidis K. Nikoloulopoulos & Dimitris Karlis

To cite this article: Aristidis K. Nikoloulopoulos & Dimitris Karlis (2010) Regression in a copula model for bivariate count data, Journal of Applied Statistics, 37:9, 1555-1568, DOI: 10.1080/02664760903093591

To link to this article: https://doi.org/10.1080/02664760903093591

Published online: 06 Sep 2010.

Submit your article to this journal

Article views: 439

Citing articles: 17 View citing articles

Journal of Applied Statistics
Vol. 37, No. 9, September 2010, 1555–1568

Regression in a copula model for bivariate count data

Aristidis K. Nikoloulopoulos[a]* and Dimitris Karlis[b]

[a] *School of Computing Sciences, University of East Anglia, Norwich, NR47TJ, UK;* [b] *Department of Statistics, Athens University of Economics and Business, 76 Patission Str., 10434 Athens, Greece*

(*Received 30 January 2008; final version received 4 June 2009*)

In many cases of modeling bivariate count data, the interest lies on studying the association rather than the marginal properties. We form a flexible regression copula-based model where covariates are used not only for the marginal but also for the copula parameters. Since copula measures the association, the use of covariates in its parameters allow for direct modeling of association. A real-data application related to transaction market basket data is used. Our goal is to refine and understand whether the association between the number of purchases of certain product categories depends on particular demographic customers' characteristics. Such information is important for decision making for marketing purposes.

Keywords: dependence modeling; Kendall's tau; covariate function; negative binomial distribution

1. Introduction

Bivariate and multivariate count data appear in a wide range of fields like epidemiology (e.g. different types of a disease, or a disease in different areas), marketing (e.g. purchases of different products), environmetrics (e.g. different kinds of plantation, etc.), criminology (e.g. different types of crimes), industrial statistics (e.g. different type of faults in a system), sports statistics (e.g. number of goals scored by each team), among others, where incidences of several related events are counted.

The existing literature contains a series of models which generalize standard univariate models. Examples are the bivariate Poisson distribution, described in detail by Kocherlakota and Kocherlakota [16] and its generalizations, based usually on mixtures as described in Aitchinson and Ho [1], Winkelmann [31], Chib and Winkelmann [5], Karlis and Xekalaki [15]. However, certain limitations are as follows: (a) they are not convenient for creating flexible models with various univariate marginal distributions and (b) since the correlation structure comes from a continuous multivariate mixing distribution, the possible choices are very limited and perhaps they lead

*Corresponding author. Email: A.Nikoloulopoulos@uea.ac.uk

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760903093591
http://www.informaworld.com

to very specific models. Some other models that have been used like the conditional model of Berkhout and Plug [2] also suffer from the difficulty of generalization to other families of marginal distributions. Thus, it seems that there is a lack of models appropriate for bivariate counts with flexible structure, for example, allowing negative dependence or models with arbitrary margins.

For this reason, we proceed by considering the use of copula functions. It is evident that the literature for copulas for count data is sparse. By definition [12,24], a bivariate copula $C(u_1, u_2)$ is a cumulative distribution function (cdf) with uniform marginals. If $F_1(y_1)$, $F_2(y_2)$ are the cdfs of the random variables Y_1, Y_2, then $C(F_1(y_1), F_2(y_2))$ is a bivariate distribution for (Y_1, Y_2) with marginal distributions F_1, F_2. Conversely, if H is a bivariate cdf with univariate marginal cdfs F_1, F_2, then there exists a bivariate copula C such for all (y_1, y_2), $H(y_1, y_2) = C(F(y_1), G(y_2))$. If F_1, F_2 are continuous, then C is unique; otherwise, there are many possible copulas as emphasized by Genest and Nešlehová [9], but all of these coincide on the closure of Range(F_1) × Range(F_2), where Range(F) denotes the range of F. The above result, known as Sklar's theorem, indicates the way that multivariate cdfs and their univariate cdfs can be connected, resulting in very flexible models by joining arbitrary univariate marginal distributions. Although the bivariate density can be derived from partial derivatives for the continuous case, for discrete data the probability mass function h is obtained using finite differences of the copula representation of H [29],

$$\begin{aligned} h(y_1, y_2) = {} & C(F_1(y_1), F_2(y_2)) - C(F_1(y_1 - 1), F_2(y_2)) \\ & - C(F_1(y_1), F_2(y_2 - 1)) + C(F_1(y_1 - 1), F_2(y_2 - 1)). \end{aligned} \tag{1}$$

There are few papers for modeling bivariate count data using copula functions. The authors mainly use comprehensive families of bivariate copulas (e.g. Frank and normal), which allow both positive and negative dependence among random variables. Lee [19] and Cameron *et al.* [4] used the Frank copula to model bivariate count data, while Van Ophem [30], Song [29], and Lee [20] used a normal copula-based model. In Escarela *et al.* [7] a Gumbel copula transition regression model was applied to time series count data.

In all the above papers, the copula parameter was taken unconditional on any covariate, while each arbitrary marginal was specified conditional on covariates. For continuous data, Lambert and Vandenhende [17] introduced covariates for the normal copula parameter, but therein marginal properties do not influence dependence. This is not true for discrete random variables because the probability of a tie is positive.

The aim of the present paper is to measure the effect of covariates on dependence structure by building a fully parametric copula-based model and allowing the specification of the dependence parameter through covariates and a link function. Note also that the existing papers for modeling multivariate count data are not helpful to measure the effect of covariates on dependence. Thus, we aim at contributing in this area by creating models for bivariate (multivariate) counts that allow to model explicitly the dependence structure, conditional on the covariate information.

The remainder of the paper proceeds as follow: In Section 2, a copula-based formula for Kendall's tau for discrete data is given. In Section 3, we introduce the bivariate copula regression model and explore the ways that covariates can affect dependence. In Section 4, we apply the proposed methods to transactional market basket data, while Section 5 concludes the paper.

2. Kendall's tau for discrete data

It is important to note that the number of different copula families proposed in the literature is quite large. Special classes of copula families such as the Archimedean, extreme value, and elliptical can be considered; [11]. In Table 1, members of these classes are listed and in particular those

Table 1. Parametric families of copulas and their link functions: note that $\bar{u}_j = 1 - u_j$ and $\tilde{u}_j = -\log u_j, \quad j = 1, 2$.

Family	$C(u_1, u_2; \theta)$	$\theta \in$	$s(\theta)$
Frank	$-\dfrac{1}{\theta}\log\left[1 + \dfrac{(\mathrm{e}^{-\theta u_1} - 1)(\mathrm{e}^{-\theta u_2} - 1)}{\mathrm{e}^{-\theta} - 1}\right]$	$(-\infty, \infty)/\{0\}$	θ
Galambos	$u_1 u_2 \mathrm{e}^{(\tilde{u}_1^{-\theta} + \tilde{u}_2^{-\theta})^{-1/\theta}}$	$[0, \infty)$	$\log\theta$
Gumbel	$\mathrm{e}^{-(\tilde{u}_1^{\theta} + \tilde{u}_2^{\theta})^{1/\theta}}$	$[1, \infty)$	$\log[\theta - 1]$
Joe	$1 - (\bar{u}_1^{\theta} + \bar{u}_2^{\theta} - \bar{u}_1^{\theta}\bar{u}_2^{\theta})^{1/\theta}$	$[1, \infty)$	$\log[\theta - 1]$
Mardia–Takahasi	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$(0, \infty)$	$\log\theta$
Normal[a]	$\Phi_\theta\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right)$	[−1,1]	$\log\left[\dfrac{1+\theta}{1-\theta}\right]$

Notes: [a] Φ, N(0,1) cdf; Φ^{-1}, functional inverse of Φ; Φ_θ, bivariate standard normal cdf with correlation θ.

which we will use in the paper. One can read from the table that copula parameters have different range, thus in order to make them comparable and interpretable we will use the copula-based Kendall's tau association.

Denuit and Lambert [6], Mesfioui and Tajar [21], and Nešlehová [25] studied the behavior of Kendall's tau applied to discrete data. In the discrete case, it is no longer distribution free and has also a range narrower than $[-1, 1]$. In the following lemma, we provide an alternative formula for Kendall's tau when the random variables are discrete. For normalized versions one can refer to Goodman and Kruskal [10] and Nešlehová [25]. This formula helps to see clearly that in the discrete case the marginals do affect dependence measures.

Lemma 2.1 *Let Y_1, Y_2 be integer-valued discrete random variables whose joint distribution is H, with marginal cdfs F_j, pmfs f_j, $j = 1, 2$, and copula C. Then the population version of Kendall's tau for Y_1 and Y_2 is given by*

$$\tau(Y_1, Y_2) = \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} h(y_1, y_2)\{4C(F_1(y_1 - 1), F_2(y_2 - 1)) - h(y_1, y_2)\} + \sum_{y_1=0}^{\infty} f_1^2(y_1) + \sum_{y_2=0}^{\infty} f_2^2(y_2) - 1, \tag{2}$$

where h is the joint pmf given in Equation (1).

Proof Let (Y_{11}, Y_{21}) and (Y_{12}, Y_{22}) denote independent copies of the vector (Y_1, Y_2). When Y_1 and Y_2 are discrete random variables,

$$P(\text{concordance}) + P(\text{discordance}) + P(\text{tie}) = 1,$$

so that we have

$$\begin{aligned}\tau(Y_1, Y_2) &= P(\text{concordance}) - P(\text{discordance}) = 2P(\text{concordance}) - 1 + P(\text{tie})\\ &= 4P(Y_{12} < Y_{11}, Y_{22} < Y_{21}) - 1 + P[(Y_{11} = Y_{12}) \cup (Y_{21} = Y_{22})],\end{aligned}$$

where

$$P(Y_{12} < Y_{11}, Y_{22} < Y_{21}) = \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} P(Y_{12} < y_1, Y_{22} < y_2)P(Y_{11} = y_1, Y_{22} = y_2)$$

$$= \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} P(Y_{12} \le y_1 - 1, Y_{22} \le y_2 - 1)P(Y_{11} = y_1, Y_{22} = y_2)$$

$$= \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} C(F_1(y_1 - 1), F_2(y_2 - 1))h(y_1, y_2)$$

and

$$P[(Y_{11} = Y_{12}) \cup (Y_{21} = Y_{22})] = P(Y_{11} = Y_{12}) + P(Y_{21} = Y_{22}) - P(Y_{11} = Y_{12}, Y_{21} = Y_{22}),$$

where

$$P(Y_{i1} = Y_{i2}) = \sum_{y=0}^{\infty} P[(Y_{i1} = y) \cap (Y_{i2} = y)] = \sum_{y=0}^{\infty} P(Y_{i1} = y)P(Y_{i2} = y)$$

$$= \sum_{y=0}^{\infty} f_i^2(y), \quad i = 1, 2$$

and

$$P(Y_{11} = Y_{12}, Y_{21} = Y_{22}) = \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} P[(Y_{11} = y_1, Y_{21} = y_2) \cap (Y_{12} = y_1, Y_{22} = y_2)]$$

$$= \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} P(Y_{11} = y_1, Y_{21} = y_2)P(Y_{12} = y_1, Y_{22} = y_2)$$

$$= \sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty} h^2(y_1, y_2). \qquad \blacksquare$$

To help the reader, in Figure 1 we have plotted Kendall's tau values for some of the copula families, that listed in Table 1, letting the copula parameter θ to vary and using Poisson marginal distributions with varying mean parameter μ equal for the two margins. One can see that for small μ Kendall's tau cannot attain large values, and different values of μ lead to different values of Kendall's tau; this can be interpreted due to the larger probability of a tie for small marginal means. For a value of μ greater than 10, the effect is negligible and the value of Kendall's tau has almost stabilized. Note that as μ tends to infinity, the upper bound of Kendall's tau is 1, due to the fact that the probability of a tie goes to 0.

3. The copula-based regression model

Consider a bivariate copula-based parametric model for the count responses Y_1 and Y_2 with distribution function H provided by the copula representation,

$$H(y_1, y_2; \alpha_1, \alpha_2, \theta) = C(F_1(y_1; \alpha_1), F_2(y_2; \alpha_2); \theta), \tag{3}$$

where F_1 and F_2 are the marginal distributions, with parameter vectors α_1 and α_2 and θ is the copula parameter.

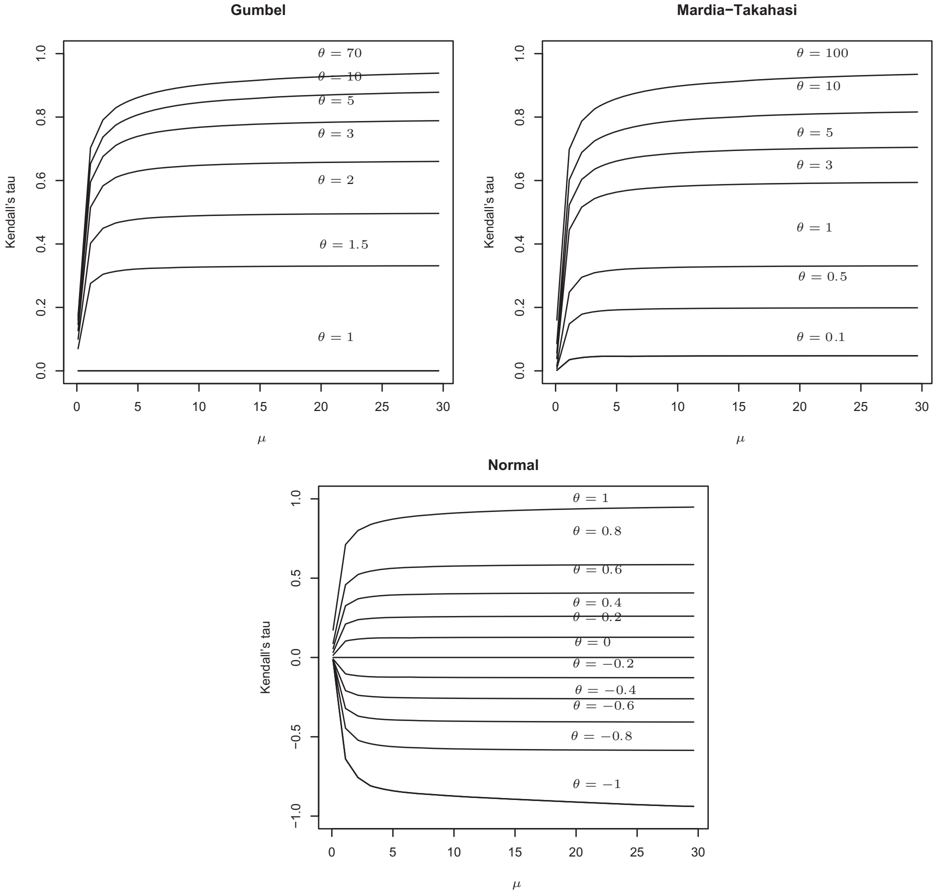

Figure 1. Kendall tau values computed using one-parameter copulas for a grid of parameter value for each copula (different lines) and Poisson marginal distributions with the same parameter μ up to 30, higher curves corresponding to higher values of the copula parameter.

3.1 *Covariates and association*

Consider the case when one wants to examine the effect of covariate information on the dependence structure. Suppose the data are $(y_{ij}, x_{ij}),\ i = 1, \dots, n,\ j = 1, 2$, where i is an index for individuals, j an index for the count responses, and x_{ij} a vector of covariates for the ith individual associated to the jth count response. One can easily introduce covariates x_{ij} on the copula-based parametric model by assuming that the univariate marginal model is $y_{ij} \sim F_j(\cdot; \alpha_{ij})$, with $\alpha_{ij} = (\mu_{ij} = g(\beta_j^{\mathrm{T}} x_{ij}), \gamma_j)$ where μ_{ij} denotes the mean parametrized by a suitable link function $g(\cdot)$ to accommodate the covariates, β_j the vector of the regression coefficients, and γ_j the vector of marginal parameters that does not depend on covariates. As mentioned earlier, marginal parameters also influence measures of dependence such as Kendall's tau. Thus, by such an approach one can measure some effect of the covariates to the Kendall's tau.

Furthermore, one can introduce a regression part for the copula parameter θ, meaning that θ can be specified conditional on the covariates with corresponding parameter vector b using an appropriate covariate function $s(\cdot)$ on θ, $s(\theta_i) = b^{\mathrm{T}} x_i$. In Table 1, the last column lists the possible covariate functions, depending on the range of the assumed copula parameter (which is not common for all copulas). These transformations are strictly increasing functions of θ and tend to $+\infty$ ($-\infty$) when θ tends to the upper (lower) bound of its range. Such an approach directly relates the copula parameter to covariates, while in the former approach the covariates affect the dependence only indirectly. The covariates in the three parts of the model can be different.

From the above discussion it is obvious that there are three possible ways to introduce covariate information related to the dependence in our model: by placing the covariate information (a) in marginal parameters, (b) in copula parameter and (c) both in marginal and copula parameters. The effect and the dependence structure implied by each one of the three approaches is different. In the sequel, these approaches will be explored when count data are involved.

3.2 *Estimation*

For estimation, we will use the standard maximum-likelihood (ML) method. The log-likelihood to be maximized is given by

$$L(\beta_1, \gamma_1, \beta_2, \gamma_2, b) = \sum_{i=1}^{n} \log h(y_{i1}, y_{i2}; \beta_1, \gamma_1, \beta_2, \gamma_2, b), \tag{4}$$

where h is the joint pmf in Equation (1). Numerical optimization is needed for deriving the ML estimates. Various optimization methods such as the quasi-Newton method [22] and the robust Nelder–Mead algorithm [23] can be used. These minimization methods require only the objective function, that is, the negative log-likelihood, while the gradients are computed numerically and the Hessian matrix of the second-order derivatives is updated in each iteration. Initial estimates for the ML estimates can be provided by the method of inference functions for margins proposed and studied by Joe [12,13]. Assuming that the usual regularity conditions [28] for asymptotic ML theory hold for the bivariate model as well as for its margins, we have that ML estimates are asymptotically normal. Therefore, one can build Wald tests to statistically judge the effect of any covariate in the dependence structure.

3.3 *Simulated evidence*

Consider the copula-based parametric model in Equation (3), where $C(\cdot; b^{\mathrm{T}} x_i)$ is the Frank copula and $F_1(\cdot; \alpha_1) = F_2(\cdot; \alpha_2) = F(\cdot; \mu_i = \exp\{\beta^{\mathrm{T}} x_i\})$ is the Poisson distribution. Both parts of the model are specified conditional on the same covariate x.

In order to visualize the effect of the covariate values on Kendall's tau, we created Figure 2 where Kendall's tau is plotted against x. In the left panel of Figure 2, the parameter values are $\beta_0 = 0$, $\beta_1 = 1$, $b_0 = 5$, $b_1 = 1$, and $x_i \in [-4, 4]$. The Kendall's tau values are calculated by plugging in the assumed parameters in Equation (2). Note here that the marginal parameter μ_i is kept smaller than 10, thus the marginals also affect dependence. The way the information from x relates to Kendall's tau varies according to whether the covariate is used (a) only for the marginals, (b) only for the copula parameter, and (c) both in the marginals and the copula parameter. Cases (a) and (b) are derived by setting $x = 0$ in the covariate function of F and C, respectively. One can see that all cases are close together for x close to 0, and they differ as x moves away from 0, implying that in this case the dependence structure implied by each model differs. In the right panel of Figure 2, the parameter values are $\beta_0 = 3$, $\beta_1 = 0.1$, $b_0 = 5$, $b_1 = 1$, and $x_i \in [-4, 4]$. One can see here that for large marginal parameters ($\mu_i > 10$) the dependence is not affected by them as we have already depicted in Figure 1. Kendall's tau value for Case (a) is constant and is not associated with the covariate values, which only affect marginal parameters. On the other hand, Cases (b) and (c) coincide.

We have also performed simulation experiments to explore how the same covariate relates to dependence when it is used correctly or incorrectly in marginal and/or the copula. We simulated

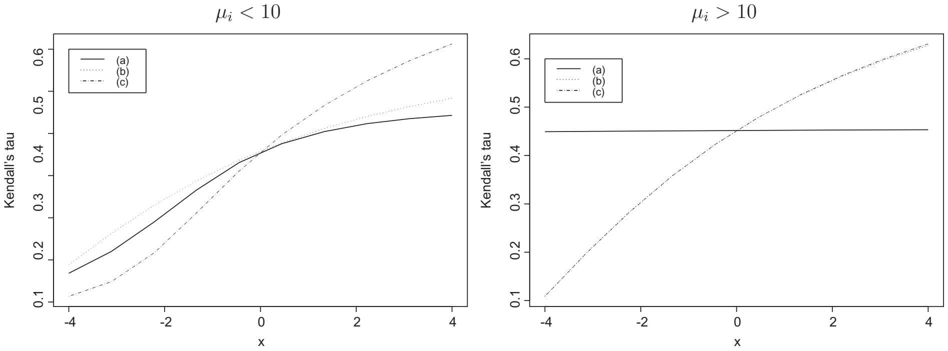

Figure 2. Kendall's tau for count data as a function of the covariate x for the three different cases: (a) covariate x influences only the marginal probabilities, (b) influences only the copula parameter and (c) both, for small ($\mu_i < 10$) and large ($\mu_i > 10$) mean values.

data from each of the three cases described above, and in the sequel we fitted all three models to see whether we can reveal the dependence structure implied. Some typical results are shown in Table 2 using the Frank copula model. The algorithm described in Genest [8] was used to simulate from the Frank copula and then the inversion method to simulate from the marginals.

The data have been drawn in different ways as described in the first column (Data), while ML estimation performed using the models listed in the second column (Model). Average estimators

Table 2. Simulation examples to explore how the same covariates affect the dependence on count data when they are placed correct or incorrect in marginal parameters and/or the Frank copula parameter.

Data	Model	Marginal parameters				Copula parameters	
$\theta_i = 5 + x_i$	I	μ_1 (se)		μ_2 (se)		b_0 (se)	b_1 (se)
$\mu_1 = 2$		2.01 (0.04)		3.01 (0.05)		5.03 (0.26)	1.02 (0.26)
$\mu_2 = 3$	II	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	θ (se)	
$y_j \sim$ Poisson(μ_j)		0.7(0.02)	0.00 (0.02)	1.1 (0.02)	0.00 (0.02)	4.92 (0.25)	
$x_i \sim N(0, 1)$	III	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	b_0 (se)	b_1 (se)
		0.7 (0.02)	0.00 (0.02)	1.1 (0.02)	0.00 (0.02)	5.03 (0.26)	1.02 (0.26)
$\theta = 5$	I	μ_1 (se)		μ_2 (se)		b_0 (se)	b_1 (se)
$\mu_{i1} = \exp\{-0.5x_i\}$		1.13 (0.04)		1.14 (0.04)		1.51 (0.2)	0.03 (0.21)
$\mu_{i2} = \exp\{0.5x_i\}$	II	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	θ (se)	
$y_{ij} \sim$ Poisson(μ_{ij})		0 (0.03)	−0.51 (0.03)	0 (0.03)	0.5 (0.03)	5.03 (0.31)	
$x_i \sim N(0, 1)$	III	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	b_0 (se)	b_1 (se)
		0 (0.03)	−0.51 (0.03)	0 (0.03)	0.5 (0.03)	5.04 (0.31)	−0.01 (0.34)
$\theta_i = 5 + x_i$	I	μ_1 (se)		μ_2 (se)		b_0 (se)	b_1 (se)
$\mu_{i1} = \exp\{0.5x_i\}$		1.16 (0.04)		1.12 (0.04)		1.52 (0.2)	0.22 (0.21)
$\mu_{i2} = \exp\{-0.5x_i\}$	II	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	θ (se)	
$y_{ij} \sim$ Poisson(μ_{ij})		0.00 (0.03)	0.50 (0.03)	0.00 (0.03)	−0.50 (0.03)	4.93 (0.3)	
$x_i \sim N(0, 1)$	III	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	b_0 (se)	b_1 (se)
		0.00 (0.03)	0.50 (0.03)	0.00 (0.03)	−0.50 (0.03)	5.03 (0.31)	0.94 (0.33)
$\theta = 5$	I	μ_1 (se)		μ_2 (se)		b_0 (se)	b_1 (se)
$\mu_1 = 2$		2.00 (0.04)		3.00 (0.05)		5.03 (0.26)	−0.01 (0.26)
$\mu_2 = 3$	II	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	θ (se)	
$y_j \sim$ Poisson(μ_j)		0.69 (0.02)	0.00 (0.02)	1.1 (0.02)	0.00 (0.02)	5.03 (0.26)	
$x_i \sim N(0, 1)$	III	β_{01} (se)	β_{11} (se)	β_{02} (se)	β_{12} (se)	b_0 (se)	b_1 (se)
		0.69 (0.02)	0.00 (0.02)	1.10 (0.02)	0.00 (0.02)	5.03 (0.26)	−0.02 (0.26)

Note: (I, $\theta_i = b_0 + b_1 x_i$, μ_j, $j = 1, 2$, without covariates; II, θ without covariates, $\mu_{ij} = \exp\{\beta_{0j} + \beta_{1j} x_i\}$; III, $\theta_i = b_0 + b_1 x_i$, $\mu_{ij} = \exp\{\beta_{0j} + \beta_{1j} x_i\}$.)

and standard errors are reported based on 100 datasets of sample size ($n = 1000$) from each example. The conclusion is that a full model including the same covariate on marginal and copula parameters describes the data satisfactory in all cases, while if we omit the covariate in some part then we may fail to describe the data structure. One can see that the full model performs better for all the cases.

As a conclusion, we recommend the use of the covariates in both the marginals and the copula parameters in order to describe the effect of a covariate in the dependence, as in the other cases we may ignore a substantial part of the structure implied by the covariate information.

4. Application to bivariate count data

4.1 *Data description*

Transactional market basket data provide excellent opportunities for a retailer to segment the customer population into different groups based on differences in their purchasing behavior. They reflect the frequency of purchases of products or product categories within the retail store and, as a result, they are extremely useful for modeling consumer purchase behavior. Moreover, they reflect the dependencies that exist between purchases made in different product categories.

We used the scanner data described in Brijs *et al*. [3], which refer to the number of purchases of food y_{i1} and non-food y_{i2} products of the ith loyalty card holder of a large super market. As food products, we refer to products used as food, while non-food category contains all the rest. The dependencies between these two categories are important for a retailer as they have different profit margin and hence identifying the relationships can help improving the retailer strategy. Our goal is to refine and understand whether these associations depend on particular demographic customers' characteristics in order to support marketing decisions.

A subset of the entire data base of size $n = 2580$ for a given time period was selected. By examining the data we found that they have overdispersion relative to the simple Poisson model, namely the overdispersion is 20.14 for the food data and 5.43 for the non-food data. In Figure 3, one can also see the large tails present in our data, as well as the dependence between food and non-food data.

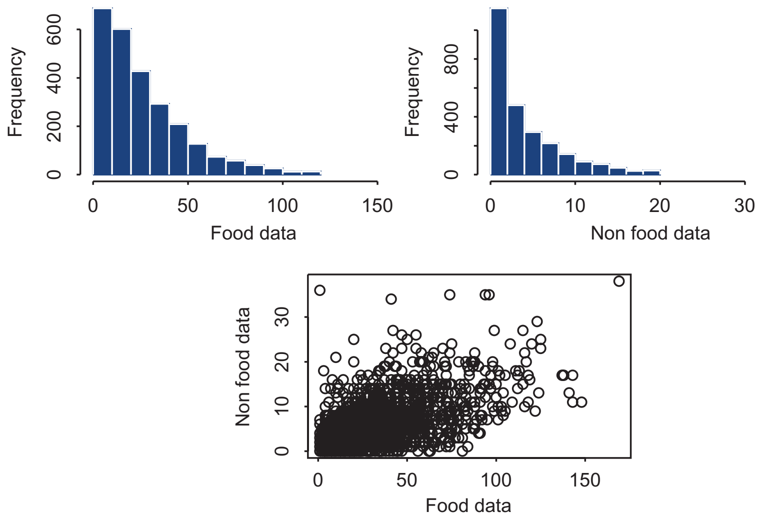

Figure 3. The purchase data.

The available covariate information refers to the customers if they use their car to go for shopping in the supermarket, if they have pets, if they have a freezer, if they have a microwave, and if they have a garden in their home. Moreover, the number of members of the family belonging to four different age subcategories: (a) 0–18 years, (b) 18–45 years, (c) 45–65 years, and (d) <65 years, was recorded to account for the household composition.

4.2 *The fitted models*

Before choosing the appropriate copula family to model the dependence among purchases, the univariate marginal distributions must be selected. For our data, the negative binomial model is considered, allowing for the large over-dispersion found in the data [18]. For each observation the cdf of each marginal is

$$F_j(y_{ij}|x_{ij},\beta_j) = \sum_{k=0}^{y_{ij}} \frac{\Gamma(\gamma_j + k)}{\Gamma(\gamma_j)\Gamma(k+1)} \frac{\mu_{ij}^k \gamma_j^{\gamma_j}}{(\mu_{ij}+\gamma_j)^{\gamma_j+k}}, \quad i = 1,\ldots,2580, \quad j = 1,2, \tag{5}$$

where $E(Y_{ij}) = \mu_{ij} = \exp\{\beta_j^{\mathrm{T}} x_{ij}\}$ and $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}^2/\gamma_j$.

Once the margins are specified, an appropriate copula function must be selected. Herein, we used six one-parameter copula families, namely Frank, Galambos, Gumbel, Mardia–Takahasi (M-T), and normal (Table 1). In the present application the marginal parameters were quite large, thus Kendall's tau is only slightly associated with marginal parameters. So, the only way that we can measure the effect of covariates on Kendall's tau is by specifying the copula parameter conditional on covariates, see Section 3. To do so, we need to specify a link function that allows to connect the covariates to the copula parameter. Such covariate functions can be read in the last column of Table 1. Note that all the available covariates were considered for the univariate marginal and the copula parameters according to Section 3.3. Table 3 contains estimates, standard errors (produced by the inversion of the Hessian), and the log-likelihood for each model.

Goodness-of-fit tests for copula-based models have found increasing popularity in recent days, see, for example, Nikoloulopoulos and Karlis [26] and the references therein. When working with discrete data things are easier; one can apply standard methods like χ^2 goodness-of-fit statistics. The idea is that expected frequencies can be easily derived and compared with the observed ones, based on the standard χ^2 statistic. We have followed this approach by deriving expected frequencies ($E_{\boldsymbol{y}}$) for each bivariate fitted copula model for all the possible vectors $\boldsymbol{y} = (y_1, y_2)$ along with the observed frequencies ($O_{\boldsymbol{y}}$) in order to form the χ^2 statistics, $\chi^2 = \sum_{\boldsymbol{y}} (O_{\boldsymbol{y}} - E_{\boldsymbol{y}})^2/E_{\boldsymbol{y}}$, where the sum is taken over all the possible vectors $\boldsymbol{y}$. The use of the standard asymptotic results implies some grouping, so as the expected frequencies to be larger than five. Alternatively, one may use a Monte Carlo approach. We report goodness-of-fit p-values based on the asymptotic result by grouping small expected frequencies. Note also that the sample size in our example is quite large and thus makes the rejection of the null hypothesis easier. One can read the resulted χ^2 statistics at the end of Table 3. The degrees of freedom are different for different models due to grouping. The reported p-values imply that the fit for certain models is satisfactory. The best fit, based on the log-likelihood values and the χ^2 statistics, was given by normal copula, while the second model was the Frank copula, as these two models are both reflection symmetric while the others are reflection asymmetric [12].

We applied Wald tests to detect statistical significant effects, assuming that the regularity conditions hold. We focus on the description for the copula parameters as our interest lies mostly on examining the associations. The variables that were found statistically significant for the copula parameter were car and pets, that is, the association differs between the customers with car and/or pets. Note here that these covariates are also statistically significant (or marginally statistically significant) for all copula models indicating a good agreement across different copulas.

Table 3. Estimated parameters, their standard errors (se), the log-likelihoods (ℓ) and the χ^2 statistics under the six copula models.

		Frank		Galambos		Gumbel		Joe		M-T		Normal	
		Estimated parameters	SE	Estimated parameters	SE	Estimated parameters	SE	Estimated parameters	SE	Estimated parameters	SE	Estimated parameters	SE
μ_{i1}	Intercept	3.02	0.08	3.03	0.08	3.04	0.08	3.05	0.08	3.02	0.08	3.06	0.08
	Car	−0.02	0.07	−0.02	0.07	−0.02	0.07	−0.04	0.07	0.00	0.07	−0.04	0.07
	Pets	0.08	0.03	0.09	0.04	0.09	0.04	0.08	0.04	0.09	0.03	0.08	0.03
	Club	−0.01	0.03	−0.01	0.03	−0.01	0.03	−0.01	0.03	0.00	0.03	−0.01	0.03
	Freezer	−0.07	0.06	−0.08	0.06	−0.09	0.06	−0.11	0.07	−0.07	0.06	−0.08	0.06
	Microwave	−0.08	0.04	−0.08	0.04	−0.07	0.04	−0.07	0.05	−0.08	0.04	−0.08	0.04
	Garden	0.18	0.06	0.17	0.06	0.18	0.06	0.19	0.06	0.17	0.06	0.18	0.06
	Age < 18	0.10	0.02	0.10	0.02	0.10	0.02	0.10	0.02	0.10	0.02	0.09	0.02
	$18 \leq$ Age < 45	0.07	0.02	0.07	0.02	0.07	0.02	0.07	0.02	0.07	0.02	0.07	0.02
	$45 \leq$ Age < 65	0.11	0.02	0.11	0.02	0.11	0.02	0.11	0.02	0.10	0.02	0.10	0.02
	Age ≥ 65	−0.02	0.03	−0.02	0.03	−0.02	0.03	−0.02	0.04	−0.02	0.03	−0.03	0.03
	γ_1	1.58	0.04	1.52	0.04	1.53	0.04	1.46	0.04	1.57	0.04	1.60	0.04
μ_{i2}	Intercept	1.33	0.10	1.32	0.10	1.32	0.10	1.33	0.11	1.31	0.10	1.33	0.10
	Car	−0.03	0.09	0.00	0.09	0.00	0.09	0.00	0.09	−0.02	0.09	−0.03	0.09
	Pets	0.09	0.04	0.08	0.05	0.08	0.05	0.07	0.05	0.10	0.04	0.08	0.04
	Club	−0.11	0.04	−0.09	0.04	−0.08	0.04	−0.08	0.04	−0.12	0.04	−0.11	0.04
	Freezer	−0.06	0.08	−0.07	0.08	−0.07	0.08	−0.09	0.09	−0.05	0.08	−0.08	0.08
	Microwave	−0.05	0.06	−0.06	0.06	−0.06	0.06	−0.05	0.06	−0.06	0.06	−0.04	0.06
	Garden	0.15	0.08	0.14	0.08	0.14	0.08	0.15	0.08	0.13	0.08	0.14	0.08
	Age < 18	0.16	0.03	0.15	0.03	0.14	0.03	0.14	0.03	0.17	0.03	0.15	0.03

	$18 \leq \text{Age} < 45$	0.07	0.02	0.07	0.02	0.07	0.02	0.06	0.03	0.07	0.02	0.07	0.02
	$45 \leq \text{Age} < 65$	0.04	0.03	0.04	0.03	0.04	0.03	0.03	0.03	0.06	0.03	0.04	0.03
	$\text{Age} \geq 65$	−0.14	0.04	−0.14	0.05	−0.15	0.05	−0.16	0.05	−0.13	0.04	−0.14	0.04
	γ_2	1.12	0.04	1.07	0.04	1.08	0.04	1.03	0.04	1.13	0.04	1.13	0.04
θ_i	Intercept	2.72	0.78	−0.39	0.17	−0.90	0.26	−0.59	0.29	−0.21	0.24	1.24	0.18
	Car	1.66	0.62	0.31	0.15	0.47	0.24	0.45	0.27	0.42	0.22	0.34	0.16
	Pets	0.86	0.33	0.12	0.07	0.17	0.09	0.13	0.10	0.21	0.09	0.20	0.08
	Club	0.10	0.27	0.02	0.06	0.02	0.09	0.00	0.10	0.07	0.08	0.02	0.07
	Freezer	0.64	0.33	0.05	0.12	0.08	0.17	0.03	0.19	0.16	0.16	0.07	0.14
	Microwave	−0.47	0.45	−0.09	0.08	−0.11	0.11	−0.10	0.13	−0.15	0.11	−0.12	0.10
	Garden	−0.32	0.57	0.02	0.12	0.01	0.16	0.10	0.19	−0.09	0.14	−0.03	0.13
	$\text{Age} < 18$	−0.02	0.18	0.02	0.04	0.02	0.05	0.04	0.06	−0.03	0.05	0.00	0.04
	$18 \leq \text{Age} < 45$	−0.04	0.19	0.00	0.04	0.00	0.05	0.01	0.06	−0.07	0.05	−0.07	0.04
	$45 \leq \text{Age} < 65$	0.22	0.18	0.03	0.04	0.03	0.06	0.01	0.07	0.02	0.05	0.00	0.05
	$\text{Age} \geq 65$	−0.08	0.34	−0.06	0.07	−0.08	0.09	−0.12	0.11	0.01	0.08	−0.10	0.08
	ℓ	−17075.39		−17089.98		−17088.67		−17183.05		−17208.31		−17064.93	
	χ^2	374.87		392.92		372.49		505.42		478.01		370.54	
	df	342		347		348		355		349		339	
	p-Value	0.106		0.045		0.175		0.001		0.001		0.115	

We have also fitted a more parsimonious normal copula-based model removing all the non-statistically significant covariates for the univariate and the copula parameters. In Table 4, one can read the results for this reduced model. The estimates are quite the same with the corresponding

Table 4. Reduced model using normal copula and only the statistical significant covariates.

		Estimated parameters	SE
μ_{i1}	Intercept	2.991	0.051
	Pets	0.071	0.034
	Microwave	−0.088	0.033
	Garden	0.122	0.039
	Age < 18	0.093	0.019
	$18 \le$ Age < 45	0.077	0.018
	$45 \le$ Age < 65	0.099	0.022
	Age ≥ 65	−0.024	0.032
	γ_1	1.593	0.044
μ_{i2}	Intercept	1.291	0.062
	Pets	0.076	0.043
	Club	−0.097	0.033
	Age < 18	0.161	0.025
	$18 \le$ Age < 45	0.077	0.023
	$45 \le$ Age < 65	0.054	0.029
	Age ≥ 65	−0.119	0.043
	γ_2	1.127	0.041
θ_i	Intercept	1.111	0.151
	Car	0.311	0.153
	Pets	0.179	0.072
(ℓ, χ^2, df, p-Value)		(−17070.60, 370.54, 339, 0.115)	

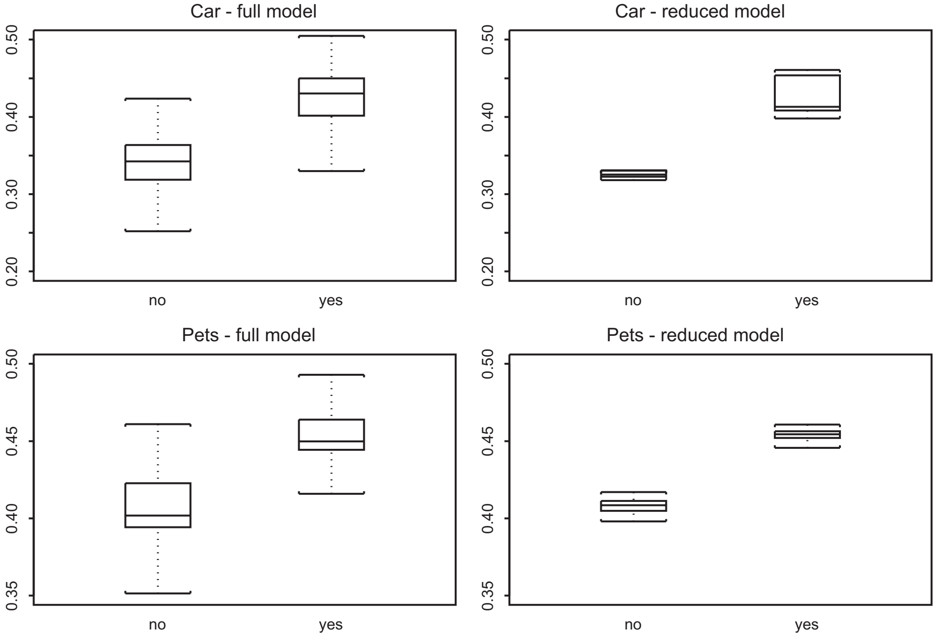

Figure 4. Box plots of Kendall’s tau using the normal copula-based model with all the covariates (full model) and with only the statistically significant ones (reduced model).

ones in Table 3, implying that the interpretation of the model is similar; the χ^2 statistic is also quite similar.

Because of the binary nature of the significant covariates, we have plotted in Figure 4 boxplots of the resulting Kendall's tau values calculated by plugging in Equation (2) the estimated parameters of the reduced normal copula-based model (right panel) and the full model with all covariates (left panel) for each observation. One can see that the mean value of Kendall's tau is very much the same, implying that both models have estimated the same average association for the different subgroups. Although the variability of the reduced model is much smaller because the noise of the non-statistical significant covariates at the copula parameter has been removed.

For customers who use their car for shopping, the dependence between food and non-food data is larger compared with those who go for shopping without their car. The same scenario appears to customers that have pets compared with them they have not. If we plan an advertisement campaign for non-food products, we expect a larger effect on the sales of food-products for those consumers with a car and/or a pet than for those consumers without these. A full examination of this from the marketing point of view is beyond our scope.

5. Concluding remarks

Copula functions are an elegant tool to model dependence in count data. Traditionally, copulas are well known in the analysis of continuous data due to their ability to totally separate the marginal from the dependence properties. This is not true for count data where there is some confounding between univariate margins and measures of associations. Introducing covariates into the dependence parameters of copulas is the main contribution of the paper.

We show that the use of covariate information can affect Kendall's tau either indirectly through the univariate margins or directly through the parameters of the copula. We examine the case when the covariates are used both in marginal and/or copula parameters, aiming to create a highly flexible model.

This modeling approach of count data can be easily extended to the multivariate case, choosing copulas that (a) provide flexible dependence, (b) have a closed form cdf, and (c) do not impose joint constraints for the vector of the copula parameters. In this end, Nikoloulopoulos and Karlis [27] used such copulas, namely the mixtures of max-id copulas defined by Joe and Hu [14], to model the associations among multivariate binary responses.

Acknowledgements

This work is part of first author's Ph.D. thesis under the supervision of the second author at the Athens University of Economics and Business. The first author would like to thank the National Scholarship Foundation of Greece for financial support. The authors would like to thank the referees and Professor Harry Joe, University of British Columbia, for comments that led to an improvement of the presentation of the paper.

References

[1] J. Aitchinson and C. Ho, *The multivariate Poisson-log normal distribution*, Biometrika 75 (1989), pp. 621–629.

[2] P. Berkhout and E. Plug, *A bivariate Poisson count data model using conditional probabilities*, Stat. Neerlandica 58 (2004), pp. 349–364.

[3] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, *Using association rules for product assortment decisions: a case study*, In: *Proceedings of the Fifth International Conference of Knowledge Discovery and Data Mining*, San Diego, CA, 1999, pp. 254–260.

[4] A.C. Cameron, T. Li, P.K. Trivedi, and D.M. Zimmer, *Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts*, Econ. J. 7(2) (2004), pp. 566–584.

[5] S. Chib and R. Winkelmann, *Markov chain Monte Carlo analysis of correlated count data*, J. Busi. Econ. Stat. 19(4) (2001), pp. 428–435.

[6] M. Denuit and P. Lambert, *Constraints on concordance measures in bivariate discrete data*, J. Multivariate Anal. 93(1) (2005), pp. 40–57.

[7] G. Escarela, R.H. Mena, and A. Castillo-Morales, *A flexible class of parametric transition regression models based on copulas: application to poliomyelitis incidence*, Stat. Methods Med. Res. 15 (2006), pp. 593–609.

[8] C. Genest, *Frank's family of bivariate distributions*, Biometrika 74(3) (1987), pp. 549–555.

[9] C. Genest and J. Nešlehová, *A primer on copulas for count data*, ASTIN Bull. 37 (2007), pp. 475–515.

[10] L. Goodman and W. Kruskal, *Measures of association for cross classifications*, J. Am. Stat. Assoc. 49 (1954), pp. 732–764.

[11] H. Joe, *Parametric families of multivariate distributions with given margins*, J. Multivariate Anal. 46 (1993), pp. 262–282.

[12] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, London, 1997.

[13] H. Joe, *Asymptotic efficiency of the two-stage estimation method for copula-based models*, J. Multivariate Anal. 94 (2005), pp. 401–419.

[14] H. Joe and T. Hu, *Multivariate distributions from mixtures of max-infinitely divisible distributions*, J. Multivariate Anal. 57(2) (1996), pp. 240–265.

[15] D. Karlis and E. Xekalaki, *Mixed Poisson distributions*, Int. Stat. Rev. 73 (2005), pp. 35–58.

[16] S. Kocherlakota and K. Kocherlakota, *Bivariate Discrete Distributions*. Marcel Dekker, New York, 1992.

[17] P. Lambert and F. Vandenhende, *A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant*, Stat. Med. 21 (2002), pp. 3197–3217.

[18] J.F. Lawless, *Negative binomial and mixed Poisson regression*, Can. J. Stat. 15 (1987), pp. 209–225.

[19] A. Lee, *Modelling rugby league data via bivariate negative binomial regression*, Aust. New Zealand J. Stat. 41 (1999), pp. 141–152.

[20] L.-F. Lee, *On the range of correlation coefficients of bivariate ordered discrete random variables*, Econo. Theory 17 (2001), pp. 247–256.

[21] M. Mesfioui and A. Tajar, *On the properties of some nonparametric concordance measures in the discrete case*, J. Nonparametric Stat. 17 (2005), pp. 541–554.

[22] J. Nash, *Compact numerical methods for computers: linear algebra and function minimisation*, 2nd ed., Hilger, New York, 1990.

[23] J. Nelder and R. Mead, A simples method for function minimization, The Computer Journal, 7 (1965), pp. 308–313.

[24] R.B. Nelsen, *An Introduction to Copulas*, Springer-Verlag, New York, 2006.

[25] J. Nešlehová, *On rank correlation measures for non-continuous random variables*, J. Multivariate Anal. 98 (2007), pp. 544–567.

[26] A.K. Nikoloulopoulos and D. Karlis, *Copula model evaluation based on parametric bootstrap*, Comput. Stat. Data Anal. 52 (2008), pp. 3342–3353.

[27] A.K. Nikoloulopoulos and D. Karlis, *Multivariate logit copula model with an application to dental data*, Stat. Med. 27 (2008), pp. 6393–6406.

[28] R.J. Serfling, *Approximation theorems of mathematical statistics*, Wiley, New York, 1980.

[29] P.X.-K. Song, *Multivariate dispersion models generated from Gaussian copula*, Scand. J. Stat. 27 (2000), pp. 305–320.

[30] H. Van Ophem, *A general method to estimate correlated discrete random variables*, Econ. Theory 15 (1999), pp. 228–237.

[31] R. Winkelmann, *Seemingly unrelated negative binomial regression*, Oxford Bull. Econ. Stat. 62 (2000), pp. 553–560.