Journal of Data Science 12(2014), 461-476

Copula-based Logistic Regression Models for Bivariate Binary Responses

Xiaohu Li [1], Linxiong Li[2*], Rui Fang[3]

[1] *University of New Orleans, Xiamen University*
[2] *University of New Orleans*
[3] *Xiamen University*

Abstract: The association between bivariate binary responses has been studied using Pearson's correlation coefficient, odds ratio, and tetrachoric correlation coefficient. This paper introduces a copula to model the association. Numerical comparisons between the proposed method and the existing methods are presented. Results show that these methods are comparative. However, the copula method has a clearer interpretation and is easier to extend to bivariate responses with three or more ordinal categories. In addition, a goodness-of-fit test for the selection of a model is performed. Applications of the method on two real data sets are also presented.

Key words: Clayton copula; Frank copula; Maximum likelihood estimation; Odds ratio; Tetrachoric correlation.

1. Introduction

When studying bivariate distributions, identifying and modeling the association structure between correlated variables is crucial. This paper focuses on modeling the dependence between the bivariate binary variables by using a copula. Due to Sklar's theorem, copulas are usually applied to model continuous data, and a discrete multivariate distribution has a copula that is uniquely determined only up to the support of the marginal distributions. Consequently this raises issues regarding the uniqueness of the copulas. We refer the reader to Genest and Neslehova [1] and Swihart, Cao, and Ciprian [2] and references therein for more discussions.

Let z denote a finite dimensional covariate vector. Suppose that two binary random variables X and Y have the following joint probability mass function,

$$\begin{cases} \mathrm{P}(X = 0, Y = 0|\mathrm{z}) = p_{00}(\mathrm{z}), & \mathrm{P}(X = 0, Y = 1|\mathrm{z}) = p_{01}(\mathrm{z}), \\ \mathrm{P}(X = 1, Y = 0|\mathrm{z}) = p_{10}(\mathrm{z}), & \mathrm{P}(X = 1, Y = 1|\mathrm{z}) = p_{11}(\mathrm{z}), \end{cases} \tag{1}$$

and the marginal probability mass functions

* Corresponding author.

$$\begin{cases} P(X = 0|\mathrm{z}) = p_0(\mathrm{z}), P(X = 1|\mathrm{z}) = p_1(\mathrm{z}), \\ P(Y = 0|z) = q_0(z), P(Y = 1|z) = q_1(z), \end{cases} \tag{2}$$

It is common in practice that observations are obtained individually from X and Y, and thus marginal probabilities p_1 and q_1 can be estimated. To suppress notations we will sometimes omit z. Commonly used estimates of the association include Pearson's correlation coefficient, Kendall's τ, and Spearman's ρ, among others. In the past decades various efforts have been made in order to model the joint probabilities based on marginal probabilities and some correlation index. We briefly describe them below. Suppose that for a given covariate vector z a logistic equation is used to model the marginal probabilities:

$$\text{logit}\,[p_1(\mathrm{z})] = \beta_1'\mathrm{z} \quad \text{and} \quad \text{logit}[q_1(\mathrm{z})] = \beta_2'\mathrm{z}.$$

- Prentice [3] used Pearson's correlation coefficient $\rho = \mathrm{Corr}(X, Y)$ to determine the joint probability

$$p_{11} = p_1 q_1 + \rho\sqrt{p_1 p_0 q_1 q_0};$$

- Dale [4] and Palmgren [5] employed the odds ratio $\varphi = \frac{p_{11}p_{00}}{p_{10}p_{01}}$ to characterize the joint probability

$$p_{11} = \begin{cases} a - \sqrt{a+b}\,, & if\ \varphi \neq 1, \\ p_1 q_1, & if\ \varphi = 1\,, \end{cases}$$

where $a = 1 + (p_1 + q_1)(\varphi - 1)$ and $b = -4\varphi(\varphi - 1)\,p_1 q_1$;

- Cessie and Houwelingen [6] considered the tetrachoric correlation and set

$$\begin{aligned} p_{11} &= P\left(\mathrm{W}_1 \le \Phi^{-1}(p_1),\ \mathrm{W}_2 \le \Phi^{-1}(q_1)\right) \\ &= \int_{-\infty}^{\Phi^{-1}(p_1)} \int_{-\infty}^{\Phi^{-1}(q_1)} \phi(t_1, t_2, \rho) dt_1 dt_2, \end{aligned}$$

where (W_1, W_2) has standard bivariate normal probability density ϕ with the correlation coefficient ρ, and Φ is the univariate standard normal distribution function. In this model the existence of latent variables is implicitly assumed.

Note that after the marginal probabilities p_1 and q_1 are obtained, the entire joint probabilities of the two binary variables are uniquely determined once p_{11} is attained. As being pointed out by many authors, the Pearson correlation coefficient is not a good dependence measure in this case because its range may be very narrow when the marginal parameters are different (McDonald [7]). So, we will only consider the odds ratio and the tetrachoric correlation models in this paper. For convenience we call ρ and φ association parameters, since they determine the joint distribution and hence the association between the binary variables.

There are discussions in the literature about the use of copulas to model the dependence and joint probability of bivariate binary variables. By our knowledge most existing studies assume that the copula is independent of covariates, which might not be true in practice. In this paper we consider that besides the marginal distributions the copula is affected by the covariates as well. The organization of the paper is as follows. Section 2 introduces copula models and related back-ground. Model selection is given in Section 3. Applications to two real data sets are presented in Section 4. Concluding remarks and comments are provided in Section 5. Detailed mathematical computation of the maximum likelihood estimation is given in Appendix.

2. Copula models

Suppose that a random vector (W_1, W_2) has joint distribution function F. Let $F1$ and $F2$ denote, respectively, the marginal distribution functions of W_1 and W_2. If there exists a function $C(u, v)$ such that

$$F(w_1, w_2) = C(F_1(w_1), F_2(w_2)), \text{ for all } w_1 \text{ and } w_2,$$

then $C(u, v)$ is called the copula of (W_1, W_2).

Since the choice of copula is independent of marginal distributions, it provides us with a convenient way to impose an association structure on marginal distributions. In the past two decades, copula has become a common tool for modeling association in biomedicine, survival analysis, financial engineering and econometrics, etc. A large number of excellent applications of copulas can be found in the literature. See for example, Shih and Louis [8], Wang and Wells [9], Wang [10] and Lakhal-Chaieb [11].

Two of the most popular copulas are Clayton copula (Clayton [12]) and Frank copula (Frank [13]). Clayton copula takes the form

$$C_\alpha(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad u, v \in (0, 1),$$

where $\alpha > -1$ and $\alpha = 0$, and Frank copula is defined by

$$C_\alpha(u, v) = -\frac{1}{\alpha} log\left(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{e^{-\alpha} - 1}\right), u, v \in (0,1),$$

where $\alpha \neq 0$. For the above two copulas α is called the association parameter. Actually, for Clayton copula, Kendall's τ can be obtained as a function of α by

$$\tau = \frac{\alpha}{\alpha + 2},$$

and for Frank copula it is

$$\tau = 4\int_0^1 \frac{e^{\alpha t}-1}{\alpha}\log\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\,\mathrm{d}t + 1.$$

It is seen that for both Clayton copula and Frank copula Kendall's τ is monotone with respect to the association parameter α, and they both can model the pair of random variables with either positive or negative association. To select a copula we in the beginning studied five commonly used copulas (Clayton, Frank, Gumbel, AMH and independent copulas) and then narrowed down to Clayton and Frank copulas based on the mathematical tractability, the interpretability in statistics, our data, and the maximum likelihood principle. So from now on we only discuss Clayton and Frank copulas. In the Appendix we use Clayton copula as an example to illustrate the procedure of maximum likelihood principle. For more on copulas and measures of association, we refer readers to Nelsen [14]. Spearman's ρ was studied as well. Similar to Kendall's τ, Spearman's ρ contains an association parameter and the ρ is strictly increasing in the parameter for both Clayton and Frank copulas. So in terms of this property Kendall's τ and Spearman's ρ are alike. However, since Kendall's τ is more convenient to use because of its explicit expression between the τ and the association parameter, we shall from now on focus on Kendall's τ alone.

Let $C_{\alpha(z)}(u,v)$ denote Clayton or Frank copula, where $u,v \in (0,1)$, and the association parameter α(z) is a function of the covariate vector z. Then, in terms of $p_1(z)$, $q_1(z)$ and $C_{\alpha(z)}$, the probability $p_{11}(z)$ is determined by

$$p_{11}(z) = C_{\alpha(z)}(p_1(z), q_1(z)). \tag{3}$$

As a consequence, other joint probabilities p_{01}, p_{10} and p_{00} can be obtained by p_1, q_1 and p_{11}.

3. The MLE and model selection

Two criteria were employed to help select a model: the maximum likelihood principle, i.e., the model that achieves the largest likelihood is selected, and a goodness-of-fit test. Remember that we shall only consider four models: odds ratio, tetrachoric correlation, Clayton copula, and Frank copula. We first formulate the likelihood. For the copula model, since the copulas under study contain one parameter α and the association between the bivariate variables is characterized by the copula, it is assumed that the copula (or the association) is affected by the covariates through α. Thus, the following three equations completely determine the joint distribution of the underlying bivariate binary variables.

$$\begin{cases} \mathrm{logit}[p_1(\mathrm{z})] = \beta_1' z, \\ \mathrm{logit}[q_1(\mathrm{z})] = \beta_2' z, \\ \ell(z) = \beta_3' z, \end{cases} \tag{4}$$

where $\ell(z)$ is a link function, and β_1, β_2 and β_3 are the corresponding coefficient vectors of z. For Clayton copula model, $\ell(z)$ = log(1 + α(z)); for Frank copula model, $\ell(z)$= α(z); for the odds ratio model $\ell(z)$ = log φ(z); and for the tetrachoric correlation model $\ell(z) = \log\frac{1+\rho(z)}{1-\rho(z)}$. It should be mentioned that the choice of the link function in each model is not unique. Note that once a sample of data is provided, the MLEs $\hat{\beta}_i$'s of $\hat{\beta}_i$'s in (4) can be obtained, and consequently the estimate of $p_1(z)$, $q_1(z)$ and $\alpha(z)$ (hence $p_{11}(z)$) can be obtained. For convenience, we shall call the first two equations in (4) *marginal probability equations* and the third *association equation.*

Suppose (X_1, Y_1), ⋯, (X_n, Y_n) are i.i.d. discrete random pairs with joint probability mass function in (1) determined by (4). Denote z_i the vector of covariates corresponding to (x_i, y_i). Then, the log-likelihood function takes the form

$$\mathcal{L}(\theta; x, y, z) = \sum_{i=1}^{n} \eta_i' \omega_i, \tag{5}$$

where

$$\theta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix}, \qquad \omega_i = \begin{pmatrix} x_i y_i \\ x_i(1-y_i) \\ (1-x_i)y_i \\ (1-x_i)(1-y_i) \end{pmatrix},$$

and

$$\eta_i = \begin{pmatrix} \log p_{11}(z_i) \\ \log\big(p_1(z_i) - p_{11}(z_i)\big) \\ \log\big(q_1(z_i) - p_{11}(z_i)\big) \\ \log\big(1 + p_{11}(z_i)\big) - p_1(z_i) - q_1(z_i) \end{pmatrix}.$$

The maximum estimator of θ is

$$\hat{\theta} = \operatorname{argmax} \mathcal{L}(\theta;\, x, y, z). \tag{6}$$

Generally, the optimization problem (6) has no analytic solution, and thus $\hat{\theta}$ has to be numerically achieved by solving the following likelihood equations

$$\sum_{i=1}^{n} [D_i(\theta)]' V_i^{-1}(\theta) \omega_i = 0, \tag{7}$$

where

$$V_i(\theta) = \begin{pmatrix} p_{11}(z_i) & 0 & 0 & 0 \\ 0 & p_{01}(z_i) & 0 & 0 \\ 0 & 0 & p_{10}(z_i) & 0 \\ 0 & 0 & 0 & p_{00}(z_i) \end{pmatrix}$$

and

$$D_i(\theta) = \begin{pmatrix} \frac{\partial p_{11}}{\partial \theta_1} & \frac{\partial p_{11}}{\partial \theta_2} & \cdots\cdots & \frac{\partial p_{11}}{\partial \theta_d} \\ \frac{\partial p_{10}}{\partial \theta_1} & \frac{\partial p_{10}}{\partial \theta_2} & \cdots\cdots & \frac{\partial p_{10}}{\partial \theta_d} \\ \frac{\partial p_{01}}{\partial \theta_1} & \frac{\partial p_{01}}{\partial \theta_2} & \cdots\cdots & \frac{\partial p_{01}}{\partial \theta_d} \\ \frac{\partial p_{00}}{\partial \theta_1} & \frac{\partial p_{00}}{\partial \theta_2} & \cdots\cdots & \frac{\partial p_{00}}{\partial \theta_d} \end{pmatrix}.$$

The estimated covariance matrix of the MLE $\hat{\theta}$ of θ is given by

$$\sum_{i=1}^{n} [D_i(\hat{\theta})]' [V_i(\hat{\theta})]^{-1} D_i(\hat{\theta}),$$

using the inverse of Fisher's information matrix evaluated at $\hat{\theta}$.

In general, the log-likelihood function and likelihood equations can be rather complicated and do not have explicit expressions. However, the likelihood equations with Clayton copula do have explicit forms, which are presented in Appendix.

In order to select a suitable model (odds ratio, tetrachoric, or a copula) to fit a data set, we used the maximum likelihood principle, i.e., the model that achieves the largest likelihood is considered to be selected. The calculation is based on the log-likelihood equation (5). In addition, a goodness-of-fit test was also used to help select a model. Since these three types of models only differ from each other in the associate equation in (4), to compare their performance, it suffices to test the hypotheses about p_{11}. For instance, for a copula model the hypotheses are

$$H_0 : p_{ij}(z) = C_{\alpha(z)}(p_1(z), q_1(z)), \text{ for all z,}$$

versus

$$H_1 : p_{ij}(z) \neq C_{\alpha(z)}(p_1(z), q_1(z)), \text{ for some z.}$$

Using the expressions of p_{11} in terms of the odds ratio and tetrachoric correlation models in Section 1, one can write similar hypotheses for the odds ratio and tetrachoric correlation models. This test can be done by implementing the classical Pearson's χ^2 test. In particular, suppose the covariate vector z is m dimensional and each component has r_i possible values, i =1, ⋯ ,m. Then, there are a total of $r_i = \prod_{i=1}^{m} r_i$ different subgroups with corresponding covariates z_k, k =1, 2,...,r. As an example, if m = 2, z_1 has r_1=2 categories (male, female), and z_2 has r_2 = 3 categories (teacher, worker, farmer), then there are r =2 × 3 = 6 subgroups with corresponding covariates z_1=(male, teacher), ⋯ , z_6=(female, farmer). With these notations the Pearson's χ^2-statistic is

$$\chi^2 = \sum_{k=1}^{r} \sum_{i,j=0,1} \frac{\left(O_{ij}(z_k) - E_{ij}(z_k)\right)^2}{E_{ij}(z_k)}$$

with 3r degrees of freedom, where O_{ij} denotes the observed frequency and E_{ij} the expected frequency, i, j =0, 1. A large value of the test statistic is evidence against the null hypothesis. We must point out that if there are continuous covariates they need to be discretized before performing the goodness-of-fit test. As will be seen in Section 4 below, in which two data sets are analyzed by using the maximum likelihood principle and the goodness-of-fit test, for Colliers data, Clayton copula is selected and for Chronic bronchial reaction data, Frank copula is preferred.

4. Two applications

In this section, we employ the copula model to analyze two real data sets of binary responses and compare the results with those obtained from the tetrachoric correlation model and the odds ratio model.

4.1 Colliers data

The data in Table 1 is from Palmgren [5] and was first presented by Ashford and Sowden [15]. This table contains frequencies of two self-reported symptoms of pneumoconiosis, breathlessness and wheezing, among working coal miners in Britain in a survey of the National Pneumoconiosis Field Trial (Fay [16]). The respondents are grouped by age and classified into nine equally spaced five-year age groups. A covariate is used for the age groups (the second column in Table 1).

Table 1: UK coalminers classified by age and self-reported symptoms

Age group in years	covariate	Breathlessness		No Breathlessness		Total
		Wheeze	No Wheeze	Wheeze	No Wheeze	
20 – 24	-4	9	7	95	1841	1952
25 – 29	-3	23	9	105	1654	1791
30 – 34	-2	54	19	177	1863	2113
35 – 39	-1	121	48	257	2357	2783
40 – 44	0	169	54	273	1778	2274
45 – 49	1	269	88	324	1712	2393
50 – 54	2	404	117	245	1324	2090
55 – 59	3	406	152	225	967	1750
60 – 64	4	372	106	132	526	1136

Table 2: Log-likelihoods for copula models

Clayton	Frank
-12858.4592	-12861.8206

than Frank copula, so between these two Clayton copula seems more suitable for this data. In addition, the likelihood values of Clayton and Frank copulas are presented in Table 2. We find that Clayton copula has a larger log-likelihood, although the difference is small. Based on the above two points, Clayton copula is preferred.

Table 3 below contains the MLE of the parameters in equations (4) for all three models under study. A score test was used to obtain the *p*-values. In this table, *Breathlessness* represents the first logistic equation for marginal probability p_1, *Wheeze* represents the second logistic equation for marginal probability q_1, and *Association* represents the correlation equation for a corresponding model.

As expected, the estimations of corresponding regression parameters of the marginal probabilities among the three models are nearly identical. Although the log-likelihood of the copula model is slightly smaller than those of the other two models, we still prefer Clayton model to the other two models for the reason that Clayton copula presents a simple, convenient connection between α (hence Kendall's τ) and covariates z by equation (4), which helps interpret the result. For example, in Table 3 the estimated slope 0.0832 of the association equation parameter of Clayton copula is significantly positive, which implies that the association between

Table 3: Parameters estimations for Colliers data

Model		Intercept (*p*-value)	Age group (*p*-value)	Log-likelihood
Tetrachoric	Breathlessness	-2.2621 (<0.0001)	0.5140 (<0.0001)	-12858.0485
	Wheeze	-1.4871 (<0.0001)	0.3253 (<0.0001)	
	Association	2.0212 (<0.0001)	0.0192 (0.18916)	
Odds ratio	Breathlessness	-2.2625 (<0.0001)	0.5145 (<0.0001)	-12858.0138
	Wheeze	-1.4878 (<0.0001)	0.3254 (<0.0001)	
	Association	3.0219 (<0.0001)	-0.1314 (<0.0001)	
Clayton copula	Breathlessness	-2.2616 (<0.0001)	0.5141 (<0.0001)	-12858.4592
	Wheeze	-1.4880 (<0.0001)	0.3252 (<0.0001)	
	Association	0.7806 (0.00015)	0.0832 (<0.0001)	

Breathlessness and *Wheeze* increases as the coal miners get older. That is, the two symptoms will have a higher probability to occur simultaneously as the age of a coal miner grows. Comparing to Clayton model, the odds ratio and tetrachoric correlation models do not have this straightforward interpretation.

To perform the goodness-of-fit test, we used the cross validation method. Let us take Clayton copula as an example to explain the procedure. The entire data set is composed of 9 age groups. There are 1952

subjects in group 1 (age 20-24). We imagine that these 1952 subjects are labeled from 1 to 1952; the second group has 1791 subjects and are labeled from 1953 to 3743 (1952+1791). Continuing doing this for all other seven age groups labels all subjects from 1 to 16122.

Step 1. Equally divide the entire data into 20 sub-groups, roughly 806 subjects in each sub-group. Subjects in each sub-group are selected randomly by tag numbers from 1 to 16122, like non-replacement lottery numbers. The first 806 random numbers constitute sub-group 1, and the second 806 constitute sub-group 2, etc. For each of the elements in sub-group 1, identify where it was in the original data. Therefore, sub-group 1 is a random subset of Table 1, the original data set. Similarly, all the other 19 sub-groups are random subsets of Table 1.

Step 2. Use one of the sub-groups to find the MLE of p_{ij} based on (3) and the other 19 sub-groups as test data sets. Thus, the χ^2 goodness-of-fit test generates 19 *p*-values. Repeat it for every sub-group.

We have done the above for 20, 35, and 50 equally divided sub-groups for Clayton copula, odds ratio, and tetrachoric correlation models. The result was that around 85% of the *p*-values were greater than 0.10 consistently across all three models and all three different number of sub-groups. Our thoughts on why 15% of the models did not pass the test include (a) the link function in equation (4) might not be appropriate and (b) more covariate(s) may be needed; "age" alone is not enough. A more precise comparison among the models has to be done by simulation studies. We will conduct the research in the future.

4.2 Chronic bronchial reaction to dust

Table 4 contains a summary of the data presented in Tutz [17]. The research was supported by the German Research Foundation. The objective of the research was focused on chronic bronchitis and dust concentration in order to determine the safe limits for the dust exposure in the workplace. All data were collected among the employees of a Munich factory (1246 employees) between 1960 and 1977. Each employee in the data had two responses: chronic bronchial reaction (yes or no) and smoking status (smoking or nonsmoking) along with two covariate variables: the concentration of dust z_1 in the workplace and the years of exposure to dust z_2. Both z1 and z2 are classified into four subgroups each. We use the midpoint of each interval (subgroup) as the covariate value.

Table 4: Chronical bronchial reaction to dust

Concentration of dust	Years of exposure	Smoking (Yes)		Smoking (No)		Total
		Yes	No	Yes	No	
0.20 – 4.00	3 – 30	71	309	16	129	525
0.20 – 4.00	31 – 66	46	148	15	47	256
4.01 – 24.0	3 – 30	70	151	9	71	301
4.01 – 24.0	31 – 66	54	72	11	27	164

The likelihood comparison in Table 5 shows a slightly larger value for Frank copula. Meanwhile, as mentioned before Frank copula has no tendency for either positive or negative

Table 5: Log-likelihoods of various copula models

Copulas	Clayton	Frank
Log-likelihood	-1352.5238	-1352.2510

association. Thus, we select Frank copula to fit this data. Table 6 below presents the maximum likelihood estimations of the regression coefficients of equations (4) and their corresponding p-values using the traditional score test.

Note that in the association equations of all three models the estimated regression coefficient of z_1 (dust concentration) is positive with a p-value around 0.05. It shows some evidence that the association between smoking and chronical bronchial reaction to dust gets stronger as the concentration of dust in the working place grows.

As far as the goodness-of-fit test is concerned, the data set was equally divided into 4, 6, and 8 sub-groups. The result was similar to that of the previous example in terms of the test rejection rate for all three models. Therefore, the conclusion is that Frank copula, odds ratio, and tetrachoric correlation models are competitive for modeling this data set.

5. Discussion and concluding remarks

A copula procedure is introduced to model the correlation of two binary variables.Although numerical examples indicate that the copula method does not outperform the existing odds ratio model or tetrachoric correlation method, the copula-based method has the advantage of extending easily to handle any discrete bivariate responses.

Pearson's correlation coefficient ρ describes only the linear relationship be-tween two variables. That is, two variables can have a value of ρ nearly zero, although they are perfectly quadratically related. By contrast, odds ratio and copula models do not suffer from such a problem, and instead they can describe all types of association. In terms of this, copula and odds ratio models are more practical than Pearson's correlation coefficient model. Although we are restricting ourselves to the Archimedean family of copulas (McNeil et al [18]), the family covers a wide range of copula selections and thus is capable of capturing various association structures. Meanwhile recall that the tetrachoric correlation model uses a pair of bivariate normally distributed variables to model the joint probability p_{11}, so it is less flexible in practice.

Table 6: Estimations of parameters

Model		Intercept (p-value)	Concentration (p-value)	Exposure (p-value)	Log-likelihood
Tetrachoric	Smoke	0.9963 (<0.0001)	0.0019 (0.46834)	0.0016 (0.42328)	-1352.3828
	Reaction	-2.4895 (<0.0001)	0.0870 (0.00011)	0.0394 (<0.0001)	
	association	0.5024 (0.18446)	0.0708 (0.05568)	-0.0096 (0.27918)	
Odds ratio	Smoke	0.9971 (<0.0001)	0.0028 (0.44979)	0.0015 (0.40394)	-1352.2431
	Reaction	-2.4884 (<0.0001)	0.0859 (<0.0001)	0.0395 (<0.0001)	
	association	0.8887 (0.02930)	0.1068 (0.04790)	-0.0191 (0.10837)	
Frank copula	Smoke	0.9969 (<0.0001)	0.0020 (0.46379)	0.0015 (0.38817)	-1352.2510
	Reaction	-2.4903 (<0.0001)	0.0870 (<0.0001)	0.0394 (<0.0001)	
	association	1.6772 (<0.0001)	0.2079 (0.04690)	-0.0352 (0.01695)	

According to McDonald [7], the MLE's of β_1 and β_2 obtained through the Pearson's correlation coefficient model or through the odds ratio model are comparative, none of the models are superior to the other. Cessie and Houwelingen [6] compared the odds ratio model and the tetrachoric correlation model using a data set in Verloove and Verwey [19], and showed similar results. Meanwhile, our calculations indicated that the MLE's of β_1 and β_2 obtained by copula models considered in this paper do not outperform the MLE's by either the odds ratio model or the tetrachoric correlation model for bivariate binary data. Therefore, for binary data all these models are competitive. However, when variables are ordinal with three or more categories, copula models can be adopted straightforwardly to characterize the association just like in the bivariate case, while the odds ratio model is less convenient in this scenario although doable by using some transformations (McCullagh and Nelder [20] and Liang et al. [21]).

When Kendall's τ or Spearman's ρ is used to measure the strength of association between two binary variables, copula models can provide an explicit expression between the covariates and Kendall's τ or Spearman's ρ. Clayton copula is a typical example. This explicit expression can help us study more properties of the variables. By contrast, the odds ratio and other models are not as convenient as copula models in this regard.

Likelihood equations in Clayton copula model

Recall that the association structure of Clayton copula is

$$C_\alpha(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad u,v \in (0,1), \quad \alpha > -1$$

Based on equations (4) we denote

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} \qquad \beta_k = \begin{pmatrix} \beta_{1,k} \\ \vdots \\ \beta_{m,k} \end{pmatrix}, \quad k = 1, 2, 3.$$

Then, d=3m and

$$\theta' = (\theta_1, \dots, \theta_d) = (\beta_{1,1}, \dots, \beta_{m,1}, \beta_{1,2}, \dots, \beta_{m,2}, \beta_{1,3}, \dots, \beta_{m,3})$$

Let $L(x) = \frac{e^x}{1+e^x}$. Then, for the log-likelihood function (5),

$$\eta_i = \begin{pmatrix} -\frac{1}{e^{\beta_3' z_i} - 1} \log\left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right) \\ \log\left(L(\beta_1' z_i) - \left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}} \right) \\ \log\left(L(\beta_2' z_i) - \left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}} \right) \\ \log\left(1 - L(\beta_1' z_i) - L(\beta_2' z_i) + \left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}} \right) \end{pmatrix}$$

The elements of the likelihood equation (7) have the explicit forms:

$$\begin{aligned} V_i(\theta) = \operatorname{diag} \Bigg\{ & \left(\left[L(\beta_1' z_i)\right]^{-e^{\beta_3' z_i}+1} + \left[L(\beta_2' z_i)\right]^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}}, \\ & L(\beta_1' z_i) - \left(\left[L(\beta_1' z_i)\right]^{-e^{\beta_3' z_i}+1} + \left[L(\beta_2' z_i)\right]^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}}, \\ & L(\beta_2' z_i) - \left(\left[L(\beta_1' z_i)\right]^{-e^{\beta_3^T z_i}+1} + \left[L(\beta_2' z_i)\right]^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}}, \\ & 1 - L(\beta_1' z_i) - L(\beta_2' z_i) + \left(\left[L(\beta_1' z_i)\right]^{-e^{\beta_3' z_i}+1} + \left[L(\beta_2' z_i)\right]^{-e^{\beta_3' z_i}+1} - 1 \right)^{-\frac{1}{e^{\beta_3' z_i}-1}} \Bigg\}, \end{aligned}$$

and for j=1,…,m, the components of the jth column of $D_i(\theta)$ are

$$\left(\frac{\partial p_{11}}{\partial \theta_j}\right)_i = \frac{z_{i,j}}{1+e^{\beta_1' z_i}} \left[\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right]^{-\frac{e^{\beta_3' z_i}}{e^{\beta_3' z_i}-1}} \left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1},$$

$$\left(\frac{\partial p_{10}}{\partial \theta_j}\right)_i = \frac{z_{i,j} L(\beta_1' z_i)}{1+e^{\beta_1' z_i}} \left\{ 1 - \left[L(\beta_1' z_i) \left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right)^{\frac{1}{e^{\beta_3' z_i}-1}} \right]^{-e^{\beta_3' z_i}} \right\},$$

$$\left(\frac{\partial p_{01}}{\partial \theta_j}\right)_i = \frac{-z_{i,j}}{1+e^{\beta_1' z_i}} \left[\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right]^{-\frac{e^{\beta_3' z_i}}{e^{\beta_3' z_i}-1}} \left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1},$$

$$\left(\frac{\partial p_{00}}{\partial \theta_j}\right)_i = \frac{z_{i,j} L(\beta_1' z_i)}{1+e^{\beta_1' z_i}} \left\{ \left[L(\beta_1' z_i) \left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1 \right)^{\frac{1}{e^{\beta_3' z_i}-1}} \right]^{-e^{\beta_3' z_i}} - 1 \right\},$$

for $j = m + 1,\ldots,2m$,

$$\left(\frac{\partial p_{11}}{\partial \theta_j}\right)_i = \frac{z_{i,j-m}}{1+e^{\beta_2' z_i}}\left[\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1\right]^{-\frac{e^{\beta_3' z_i}}{e^{\beta_3' z_i}-1}}\left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1},$$

$$\left(\frac{\partial p_{10}}{\partial \theta_j}\right)_i = \frac{z_{i,j-m}L(\beta_2' z_i)}{1+e^{\beta_2' z_i}}\left\{1 - \left[L(\beta_2' z_i)\left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1\right)^{\frac{1}{e^{\beta_3' z_i}-1}}\right]^{-e^{\beta_3' z_i}}\right\},$$

$$\left(\frac{\partial p_{01}}{\partial \theta_j}\right)_i = \frac{-z_{i,j-m}}{1+e^{\beta_2' z_i}}\left[\left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1\right]^{-\frac{e^{\beta_3' z_i}}{e^{\beta_3' z_i}-1}}\left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1},$$

$$\left(\frac{\partial p_{00}}{\partial \theta_j}\right)_i = \frac{z_{i,j-m}L(\beta_2' z_i)}{1+e^{\beta_2' z_i}}\left\{\left[L(\beta_2' z_i)\left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1\right)^{\frac{1}{e^{\beta_3' z_i}-1}}\right]^{-e^{\beta_3' z_i}} - 1\right\},$$

and

$$\begin{aligned}\left(\frac{\partial p_{11}}{\partial \theta_j}\right)_i = {} & \frac{e^{\beta_3' z_i} z_{i,j-2m}}{(e^{\beta_3' z_i}-1)^2}\left[\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1\right]^{\frac{1}{1-e^{\beta_3' z_i}}} \\ & \cdot\left\{\log\left(\left(L(\beta_1' z_i)\right)^{-e^{\beta_3' z_i}+1} + \left(L(\beta_2' z_i)\right)^{-e^{\beta_3' z_i}+1} - 1\right)\right. \\ & + (1-e^{\beta_3' z_i})\left[-L(\beta_1' z_i)\left(L(\beta_2' z_i)\right)^{e^{\beta_3' z_i}} + \left(L(\beta_1' z_i)\right)^{e^{\beta_3' z_i}}\left(\left(L(\beta_2' z_i)\right)^{e^{\beta_3' z_i}} - L(\beta_2' z_i)\right)\right]^{-1} \\ & \left.\cdot\left[L(\beta_1' z_i)\left(L(\beta_2' z_i)\right)^{e^{\beta_3' z_i}}\log\left(L(\beta_1' z_i)\right) + L(\beta_2' z_i)\left(L(\beta_1' z_i)\right)^{e^{\beta_3' z_i}}\log\left(L(\beta_2' z_i)\right)\right]\right\},\end{aligned}$$

$$\left(\frac{\partial p_{10}}{\partial \theta_j}\right)_i = \left(\frac{\partial p_{00}}{\partial \theta_j}\right)_i = \left(\frac{\partial p_{01}}{\partial \theta_j}\right)_i = -\left(\frac{\partial p_{11}}{\partial \theta_j}\right)_i.$$

for $j = 2m + 1,\ldots,3m$.

References

[1] Genest, C., Neslehova J. A primer on copulas for count data. *The Astin Bulletin* 2007; **37**, 475-515.

[2] Swihart, BJ., Ca'o, B., Crainiceanu, C. A united approach to modeling multivariate binary data using copulas over partitions (July 2010). Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 213.

[3] Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**, 1033-1048.

[4] Dale JR. Global cross-ratio models for bivariate discrete ordered responses. *Biometrics* 1986; **42**, 909-917.

[5] Palmgren J. Regression models for bivariate binary responses. *UW Biostatistics Working Paper Series* 1989; paper 101.

[6] Cessie S, Houwelingen J. Logistic regression for correlated binary data. *Journal of the Royal Statistical Society*: *Series C*(*Applied Statistics*) 1994; **43**, 95-108.

[7] McDonald BW. Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society*: *Series B* 1993; **55**, 391-397.

[8] Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 1995; **51**, 1384-1399.

[9] Wang W, Wells, MT. Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* 2000; **95**, 62-72.

[10] Wang A. The analysis of bivariate truncated data using the Clayton copula model. *The International Journal of Biostatistics* 2007; **3**, article 8.

[11] Lakhal-Chaieb ML. Copula inference under censoring. *Biometrika* 2010; **97**, 505-512.

[12] Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronical disease incidence. *Biometrika* 1978; **65**, 141-151.

[13] Frank MJ. On the simultaneous associativity of $F(x,y)$ and $x + y - F(x,y)$. *Aequationes Mathematicae* 1979; **19**, 194-226.

[14] Nelsen BR. *An Introduction to Copula*. Springer: New York, 2006.

[15] Ashford JR, Sowden RR. Multivariate probit analysis. *Biometrics* 1970; **26**, 535-546.

[16] Fay JWJ. The national coal board's pneumoconiosis research. *Nature* 1957; **180**, 309.

[17] Tutz G. *Regression for Categorical Data*. Cambridge University Press: New York, 2011.

[18] McNeil AJ, Frey R, Embrechts P. *Quantitative Risk Management*. Princeton University Press: New York, 2005.

[19] Verloove SP, Verwey RY. *Project on preterm and small-for-gestational age infants in the Netherlands*, thesis. University of Leiden: Leiden, 1988.

[20] McCullagh P, Nelder JA. *Generalized Linear Models*, 2nd edn. Chapman and Hall: London,1989.

[21] Liang K, Zeger S, Qaqish B. Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society: Series B* 1992; **54**, 3-40.

Received July 30, 2013; accepted October 30, 2013.

Xiaohu Li
University of New Orleans, Xiamen University
Xiamen 361005,China
xhli@lzu.edu.cn

Linxiong Li
University of New Orleans
Office: Math. Building Room249.
lli1@uno.edu

Rui Fang
Xiamen University
mathxhli@hotmail.com