Journal of Statistical Theory and Applications
Vol. **18(3)**, June 2019, *pp.* 182–197
DOI: https://doi.org/10.2991/jsta.d.190306.009; ISSN 1538-7887
https://www.atlantis-press.com/journals/jsta

Gaussian Copula–based Regression Models for the Analysis of Mixed Outcomes: An Application on Household's Utilization of Health Services Data

Z. Rezaei Ghahroodi[1,2,*], R. Aliakbari Saba[1], and T. Baghfalaki[2]

[1]*School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran*
[2]*Statistical Research and Training Center, Tehran, Iran*

ARTICLE INFO

Article History
Received 24 June 2017
Accepted 16 April 2018

Keywords
Copula models
mixed outcomes
sampling weights
marginal model

ABSTRACT

In analyzing most correlated outcomes, the popular multivariate Gaussian distribution is very restrictive and therefore dependence modeling using copulas is nowadays very common to take into account the association among mixed outcomes. In this paper, we use Gaussian copula to construct a joint distribution for three mixed discrete and continuous responses. Our approach entails specifying marginal regression models for the outcomes, and combining them via a copula to form a joint model. Closed form for likelihood function is obtained by considering sampling weights. We also obtain the likelihood function for mixed responses where one of the responses, time to event outcome, may have censored values. Some simulation studies are performed to illustrate the performance of the model. Finally, the model is applied on data involving trivariate mixed outcomes on hospitalization of individuals, based on the survey of household's utilization of health services.

1. INTRODUCTION

Many statistical applications especially in official statistics involve the collection of multivariate data comprising of a mixture of correlated discrete and continuous responses. Mixed outcomes are ubiquitous in applications and can be also found in health surveys where data may involve a patient choice of healthcare unit, her/his state of health, her/his hospital cost, her/his length of stay in hospital, result of hospitalization, and type of received services along with a number of quantitative demographic and health-related variables. Analyzing each response separately may give misleading results and multivariate modelling of such data often leads to complications in practice due to a relative lack of existence of standard models.

Factorization approach directly specify the joint distribution of variables as the product of a conditional distribution of a set of variables given other variables and a marginal distribution of the others [1].

Indirect approaches to specifying mixed outcomes joint distribution have also been studied. One approach introduces shared or correlated random effects to incorporate correlations between variables in the resulting joint model. The basic idea in this approach is to use random effects to build in correlation between mixed variables [2, 3, 4, 5, 6].

As mentioned, there is so many ways to consider the correlation structure between mixed correlated responses. When mixed responses are recorded on different scales, the copula approach is one of the best to consider the dependence structure. Therefore, a recent alternative strategy involves the use of copulas, as discussed by Sklar [7] who established theoretical basis and by Embrechts *et al.* [8] who make copulas popular in finance. Copulas incorporate the information on the dependence structure between two or more random variables. A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform and copulas are used to describe the dependence between random variables. If this dependency ignores, it may lead to wrong inference. Some of the literature on copulas focusses on the bivariate case. In this case, the approach relates an arbitrary joint distribution $F_{X,Y}$ to its corresponding univariate marginal distributions, F_X and F_Y, via a copula C as $F_{X,Y}\left(x,y\right) = C\left(F_X\left(x\right), F_Y\left(y\right);\rho\right)$, where $F_X\left(X\right)$ and $F_Y\left(Y\right)$ are, respectively, realizations of probability integral transforms $F_X\left(x\right) \sim$ uniform (0, 1) and $F_Y\left(y\right) \sim$ uniform (0, 1), and ρ is the dependency parameter measuring the dependence between marginal distributions F_X and F_Y. There were also many construction schemes for higher-dimensional copulas which consider variance–covariance structure (Σ) for taking into account the dependency among variables [9, 10, 11, 12]. The use of copulas is, however, challenging in higher dimensions, since standard multivariate copulas suffer from having flexible structures. Vine copulas

*Corresponding author. *Email:* zrezaeighahroodi@gmail.com

overcome such limitations and are able to model complex dependency patterns by benefiting from the rich variety of bivariate copulas as building blocks [13]. Joe [14] initially proposed vine copulas and in more detail Bedford and Cooke [15, 16] and Kurowicka and Cooke [17] developed them.

Different copula models are also used in this situation. Nikoloulopoulos and Karlis [18] used multivariate logit copula and de Leon and Wu [19] used copula-based regression models for a bivariate mixed discrete and continuous outcome [20–23]. Recently, Jiryaie *et al.* [24] use Gaussian copula distributions for mixed data, with application in discrimination. Also, Stober *et al.* and Zilko and Kurowicka [25, 26] used copula-based regression models for mixed discrete and continuous outcomes.

A large number of researches have concentrated their attention on the field of reliability and cost–benefit analysis (CBA) with a copula approach [27, 28]. CBA is a systematic approach to estimate the strengths and weaknesses of alternatives. The CBA is also defined as a systematic process for calculating and comparing benefits and costs of a decision, policy, or project. As an example, in our applied data set on hospitalization of individuals, based on the survey of household's utilization of health services (UHS), the benefit of increasing literacy on raising the probability of a full recovery, in relatively short time and low cost, will be illustrated.

In this paper, a closed form for the likelihood, considering three correlated random variables, is given where sampling weights are also taking into account (these weights have not been considered by others). Since applied data set in this paper is from sampling survey, we weigh the data on each member of the sample household to obtain unbiased parameter estimates. This is done in three steps: (1) using base weight which incorporates features of the complex sampling design, multistage sampling, of the UHSs, (2) weighs based on unit nonresponse, and (3) weighs based on population projection. Therefore, the weights should be used in the likelihood and simulation study. These weights somehow increase the number of observations and so more information is incorporated. Consequently, the estimation of standard errors (S.Es.) of parameter estimates would be more precise. We also illustrate the likelihood for three mixed responses where one of the responses, time to event outcome, may have censored values.

The paper is organized as follows. In Section 2, we describe the data set. In Section 3, we introduce a class of copula-based regression models for trivariate mixed outcomes, by considering sampling weights. In this section, we also consider joint modeling of three outcomes, an ordinal outcome (result of hospitalization), a time to event outcome (duration of hospitalization), and a Gaussian continuous outcome (logarithm of cost of hospitalization). Some statistical issues in analyzing mixed outcomes surveys such as sampling weights and censoring are overcame in this section. For each case, the likelihood function is given. In Section 4, some simulation studies are performed for illustration of the performance of the model. In Section 5, data on hospitalization of individuals from survey of household's UHSs are used to illustrate our methodology. Section 6 gives some conclusion.

2. MOTIVATION: HOUSEHOLD'S UTILIZATION OF HEALTH SERVICES DATA

In this section, we will describe the household's UHSs data. The three selected response variables, extracted from these data, are duration and costs of hospitalization as continuous variables and result of hospitalization as a discrete variable. In this section, some factors that affect the behavior of responses will be descriptively examined.

The Iranian 2015 UHS survey was designed and implemented with the aim of identifying the needs of individuals and households to get health services, to be sure of availability of services and to use provided services. Designing of this survey was done regarding to some environmental, economic, social, and political factors impacted on creating health inequalities between different groups of population. The information required for calculating health equity indices and indicators in UHS were gathered through a representative sample selected from different social groups of the population. Many researches are done in different countries based on data on UHS. Garcia-Subirats *et al.* [29] investigated inequities in access to healthcare in different health systems. Also, Bastos *et al.* [30] estimate the healthcare utilization and factors influencing it in the public sector in Brazil.

In Iran, this cross-sectional household survey was implemented by Statistical Research and Training Center and the Statistical Centre of Iran **(SCI)** in cooperation with the Ministry of Health and Medical Education and the National Institute of Health Research. Sample size was 22,470 households and questionnaires were completed with face to face interview with household members. The survey has been implemented in fall of 2014 and all of the information are gathered from respondents in the time of the survey. However, the reference time of some questions such as outpatient and inpatient health services are different. Information about outpatient health service was gathered for the two last weeks before survey and information about the inpatient health service was gathered for a year from 2013 fall to survey time. In the needs and inpatient services section, the survey questionnaire was designed on the needs of individual members of the sample households to hospital (or health facility) during the fall of 2013 until the survey time and in the case of need for hospitalization, name of needs, area of creation (feeling sick or medical guidance), and how to deal with each of them were determined. For individuals who their needs led to stay in hospital (or health facility), evaluation of each inpatient was recorded by household member. For these people, in addition to individual and household information, other information such as the type of hospital, the number of hospitalizations, the waiting period, the works undertaken, the costs of hospitalization, and the result of hospitalization were registered in the questionnaires and gathered at survey time. The results show that the number of households in the sample who have been hospitalized and their data was collected completely in the survey was 2486 households. In this paper, some models were fitted based on the completed sample. The characteristics for selected response variables along with factors that affect the behavior of response variables were also examined. The selected three response variables include duration of hospitalization, costs of hospitalization, and result of hospitalization.

Cost of hospitalization is a continuous variable recorded in rials and the natural logarithm of this variable is used to ensure normality of the response. Duration of hospitalization in days is the second continuous response and the result of hospitalization, as the third response, is an ordinal variable categorized as full recovery, partial recovery, and ineffective admitted. To study the behavior and the factors influencing the response variables, explanatory variables such as residence area, gender, type of hospital, literacy status, activity status, marriage status, and services are extracted from questionnaires of UHS survey and used for the analysis. The descriptive statistics of responses and also the distribution of data in different levels of each explanatory variable are given in Table 1. This table shows that 47.1% of people who leave the hospital are fully recovered. The average age of respondents who are hospitalized is 45.65 years and the mean of duration of hospitalization is 3.77 days.

Figure 1 shows histogram of the natural logarithm of the costs of hospitalization. Figure 2 shows the relationship between different covariates and logarithm of cost of hospitalization by box plots. This figure reveals that the type of hospital and service are important factors on this response. Figure 3 gives stackplots of results of hospitalization versus gender and economic activity status. Based on this figure, the subpopulation of women and men have different patterns of result of hospitalization, that is, among the men subpopulation, partial recovery has the highest percentage, but among the women subpopulation, full recovery has the highest percentage. Also, the subpopulation of employed, unemployed, and inactive people have different patterns of result of hospitalization. Among the unemployed subpopulation, partial recovery has the highest percentage.

For a primary description of the explanatory variables on duration of hospitalization, Fig. 4 shows Kaplan–Meier estimators of the survival curves of duration of hospitalization for different covariates groups by considering individual weights. Description of this figure is simple, for example, according to Fig. 4(a) males have longer duration of hospitalization than that of females. Duration of hospitalization in government hospitals is longer than that of private hospitals. Also, people with higher education have smaller duration of hospitalization than that of people who are illiterate or diploma. However, these interpretations are only correct marginally and in the presence of other covariates, using a statistical model, we may find more realistic interpretation. Figure 5 as a survival curve of duration of hospitalization in different combination level of gender and literacy shows that the lowest duration of hospitalization is for female with higher education, but for male, different levels of literacy have nearly the same duration. This means that we need to consider the interaction effect of gender and literacy.

3. JOINT MODEL FOR THREE MIXED VARIABLE OUTCOMES

In this section, copula-based regression models for the analysis of three correlated mixed outcomes are discussed. We use Gaussian copula to construct a joint distribution for the three mixed responses. In this approach, we specify marginal regression models for the outcomes, and

Table 1 | Descriptive statistics of responses and covariates.

Continuous response variables	**Level**	**No. of observation**	**Mean**
log (cost of hospitalization)	-	2486	13.0
Duration of hospitalization	-	2486	3.77
Ordinal response variable	Level	No. of observation	Percentage
Result of hospitalization	Full recovery	1171	47.1
	Partial recovery	1119	45.
	Ineffective admitted	196	7.9
Covariate variables	Level	No. of observation	Percentage
Residence area	Urban	1669	67.1
	Rural	817	32.9
Gender	Male	1002	40.3
	Female	1484	59.7
Type of hospital	Governmental	1721	69.2
	Private	491	19.7
	Others	274	11.1
Literacy status	Illiterate	707	28.4
	Diploma	1452	58.4
	Higher education	327	13.2
Activity status	Employed	608	24.5
	Unemployed	122	4.9
	Inactive	1756	70.6
Marriage status	Married	2014	81.0
	Divorced-widow	234	9.4
	Not married	238	9.6
Services	Specification	149	6.0
	Treatment	43	1.7
	Surgery	691	27.8
	Medico	844	34.0
	Rehabilitation	308	12.4
	Child birth	451	18.1

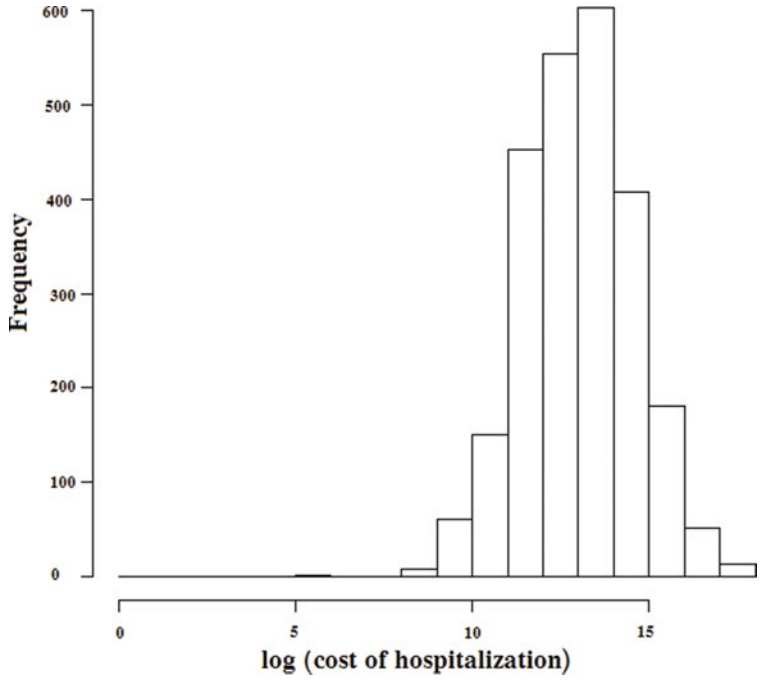

Figure 1 | Histogram of natural logarithmic of hospital costs.

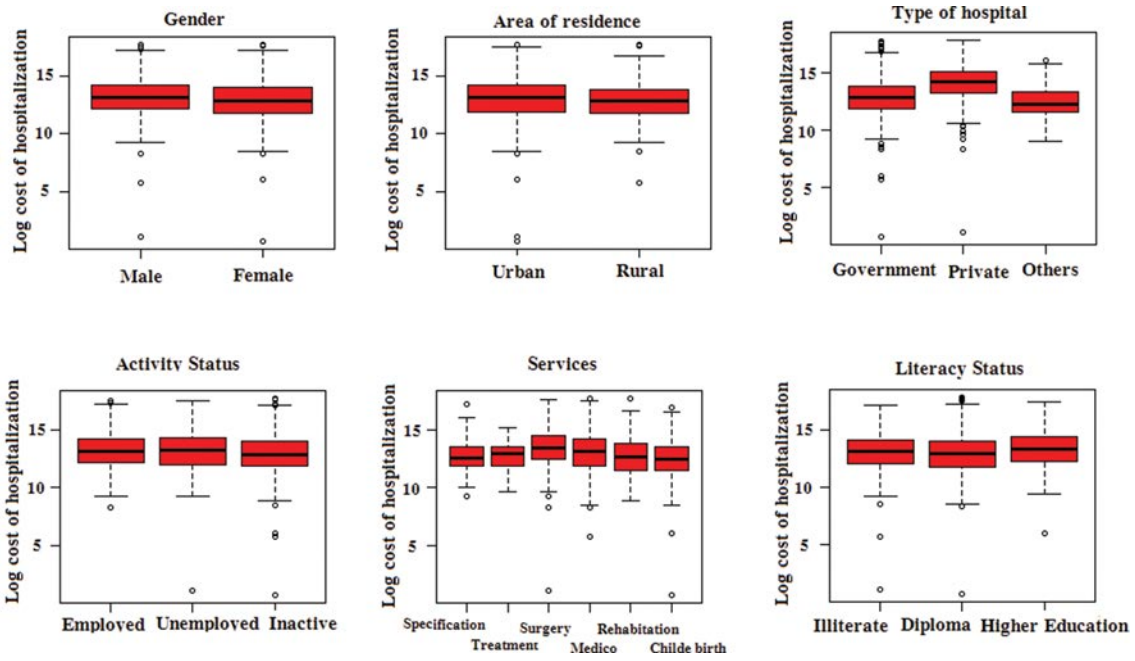

Figure 2 | The box plot of logarithm of cost of hospitalization by different covariates.

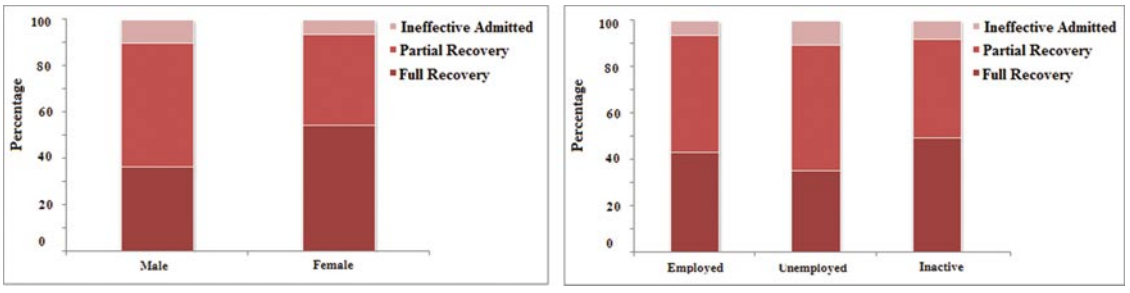

Figure 3 | The relative distribution of result of hospitalization by gender and economic activity status.

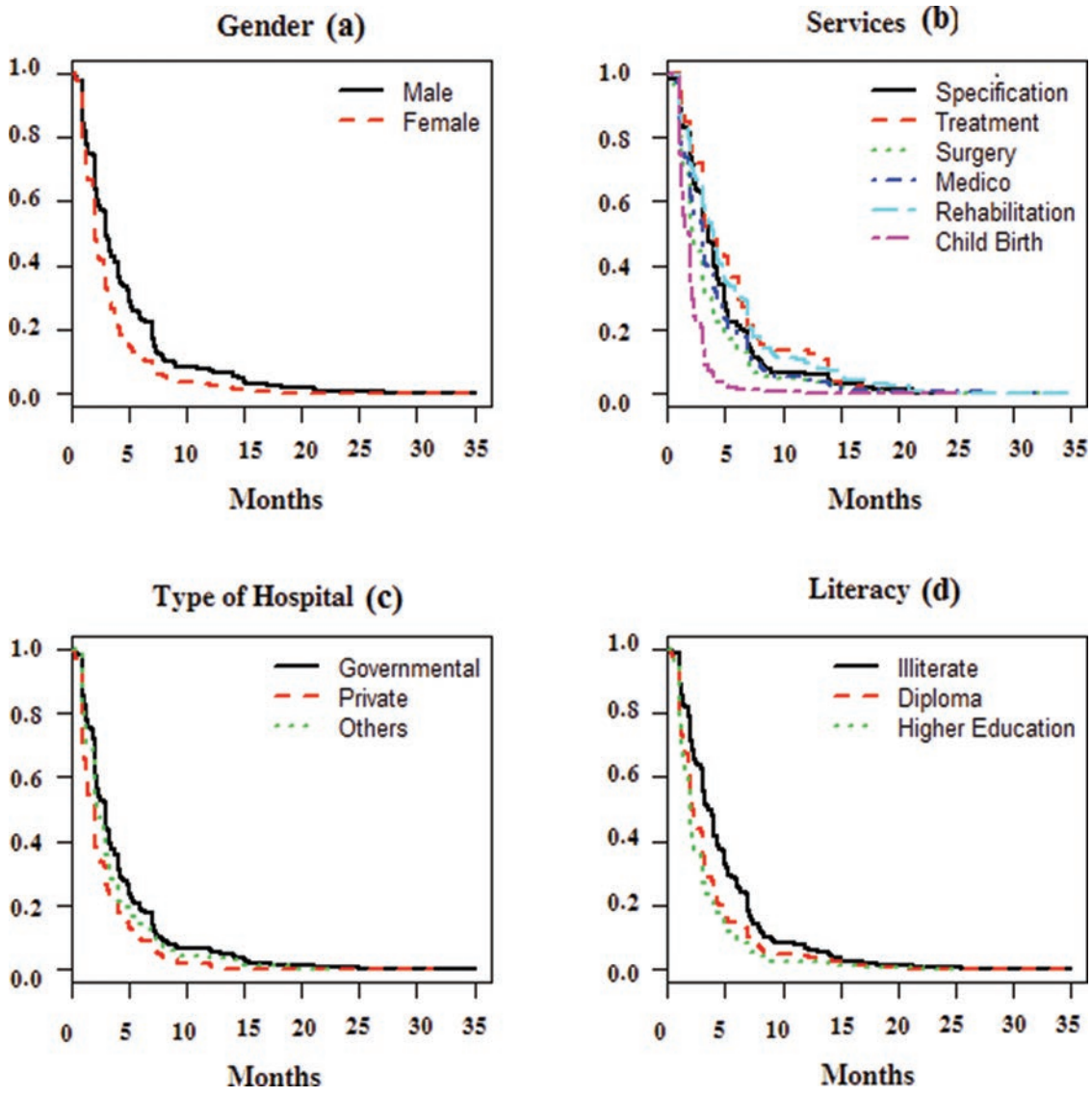

Figure 4 Survival curve of duration of hospitalization on covariates groups, (a) gender, (b) services, (c) type of hospital, (d) literacy.

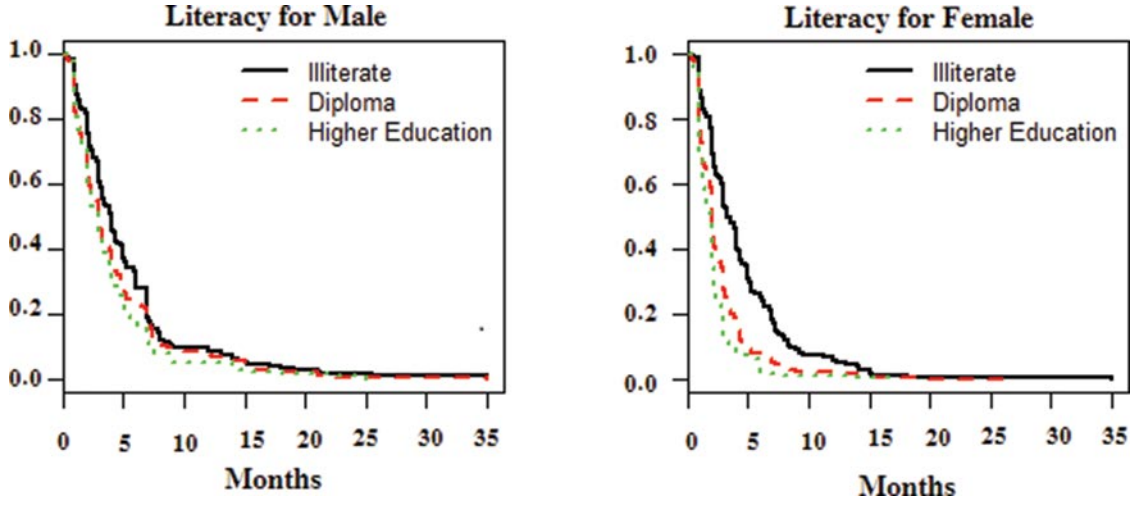

Figure 5 Survival curve of duration of hospitalization in different combination level of gender and literacy.

combine them via a copula to form a joint model. The closed form for likelihood function is also obtained by considering some sampling weights. We also obtain the likelihood function for mixed responses where one of the responses, time to event outcome, may have censored values.

Consider correlated mixed outcomes Z_i, Y_i, and T_i from each of N subjects, where Y_i is a continuous response, T_i is a time to event outcome, and Z_i is a discrete ordinal response. These outcomes may arise in a study involving patients who are admitted to a hospital, for example, the discrete outcome represents result of hospitalization in an ordinal scale, the continuous responses are the logarithm of the hospital cost and length of stay in the hospital of which the latter is a time to event response.

Assume $Z_i \sim F_{Z_i}$, $Y_i \sim F_{Y_i}$, and $T_i \sim F_{T_i}$, $i = 1, ..., N$. In addition, let Z_i have $C + 1$ distinct values, say, s_0, s_1,..., s_C, possibly representing ordinal scores. In order to use Gaussian copula, the unobserved continuous latent variable, Y_i^*, underlying the ordinal response, Z_i, should be defined. To model the joint distribution F_{Z_i,Y_i,T_i} of Z_i, Y_i, and T_i, let Y_i^*, with distribution function $F_{Y_i^*}$, be the unobserved continuous latent variable underlying Z_i, such that

$$Z_i = \begin{cases} s_0 & Y_i^* \in (-\infty, \gamma_1), \\ \vdots & \vdots \\ s_k & Y_i^* \in [\gamma_k, \gamma_{k+1}], \\ \vdots & \vdots \\ s_C & Y_i^* \in [\gamma_C, \infty]. \end{cases} \tag{1}$$

where $\gamma_0 < \gamma_1 < ... < \gamma_C < \gamma_{C+1}$, $\gamma_0 = -\infty$, $\gamma_{C+1} = \infty$, and $\gamma_1, ..., \gamma_C$ are unknown thresholds. In what follows, we use the Gaussian copula, an important copula family. This copula has been used in a variety of applications due to its flexibility and analytical tractability. We assume the joint distribution of Y_i^*, Y_i, and T_i, $F_{Y_i^*,Y_i,T_i}$, to be determined by a Gaussian copula, that is,

$$F_{Y_i^*,Y_i,T_i}\left(y_i^*, y_i, t_i\right) = \Phi_3\left[\Phi^{-1}\left(F_{Y_i^*}\left(y_i^*\right)\right), \Phi^{-1}\left(F_{Y_i}\left(y_i\right)\right), \Phi^{-1}\left(F_{T_i}\left(t_i\right)\right); R\right] \tag{2}$$

where Φ is the standard normal distribution, $\Phi_3(.,.,.;R)$ is a trivariate normal distribution with zero mean and correlation matrix R. As mentioned above Z_i is an ordinal variable, we use a latent variable model for modeling variable Z_i as follows:

$$Y_i^* \sim N\left(\mathbf{x}_{zi}'\beta, 1\right).$$

As we have cut points parameters (γ_i)'s, to ensure identifiability, we have to use a latent variable model without any intercept parameter (in fact the role of intercept is played by cut points).

For continuous variable, we consider $Y_i \sim N\left(\mathbf{x}_{ij}'\beta_1, \sigma^2\right)$, and for the time to event response we have, $T_i \sim Weibull(\eta_i, r)$, so in overall, we have

$$\begin{aligned} \mu_i(\mathbf{x}_{zi}, \beta) &= \mathbf{x}_{zi}'\beta, \\ \mu_{i1}\left(\mathbf{x}_{yi}, \beta_1\right) &= \mathbf{x}_{yi}'\beta_1, \\ \mu_{i2}(\mathbf{x}_{ti}, \beta_2) &= exp\left(\mathbf{x}_{ti}'\beta_2\right) = \eta_i. \end{aligned}$$

Correlation matrix (R) of Eq. (2) is defined as follows:

$$R = \begin{bmatrix} 1 & \vdots & \rho_{zy} & \rho_{zt} \\ \cdots & & \cdots & \cdots \\ \rho_{zy} & \vdots & 1 & \rho_{yt} \\ \rho_{zt} & \vdots & \rho_{yt} & 1 \end{bmatrix} = \begin{bmatrix} 1 & R_{12} \\ R_{21} & R_{22} \end{bmatrix},$$

where $R_{12} = \left(\rho_{zy}, \rho_{zt}\right)$ and

$$R_{22} = \begin{bmatrix} 1 & \rho_{yt} \\ \rho_{yt} & 1 \end{bmatrix}.$$

Here, the marginal distributions $F_{Y_i^*}$, F_{Y_i}, and F_{T_i} are absolutely continuous distributions; β, β_1, and β_2 are vectors of regression coefficients; $\mathbf{x}_{zi}$, $\mathbf{x}_{yi}$, and $\mathbf{x}_{ti}$ are outcome-specific covariate vectors corresponding to Y_i^*, Y_i, and T_i, respectively, and μ_i, μ_{1i}, and μ_{2i} are link functions specifying how the covariates are incorporated in the marginal means.

The joint distribution of Z_i, Y_i and T_i is then given by

$$P\left(Z_i = z_i, Y_i \le y_i, T_i \le t_i\right) = \begin{cases} F_{Y_i^*,Y_i,T_i}\left(\gamma_1, y_i, t_i\right) & Z_i = s_0, \\ F_{Y_i^*,Y_i,T_i}\left(\gamma_2, y_i, t_i\right) - F_{Y_i^*,Y_i,T_i}\left(\gamma_1, y_i, t_i\right) & Z_i = s_1, \\ \vdots & \vdots \\ F_{Y_i^*,Y_i,T_i}\left(\gamma_{k+1}, y_i, t_i\right) - F_{Y_i^*,Y_i,T_i}\left(\gamma_k, y_i, t_i\right) & Z_i = s_k \\ \vdots & \vdots \\ F_{Y_i,T_i}\left(y_i, t_i\right) - F_{Y_i^*,Y_i,T_i}\left(\gamma_C, y_i, t_i\right) & Z_i = s_C. \end{cases}$$

where $F_{Y_i^*,Y_i,T_i}(.,.,.)$ is defined in Eq. (2), $\gamma_1, \gamma_2, \ldots$, and γ_C are unknown thresholds and $s_0, s_1, \ldots$, and s_C are $C+1$ distinct values of Z_i. As a demonstration of the joint distribution function of Z_i, Y_i, and T_i, the joint distribution function of Y_i^*, Y_i, and T_i for $Z_i = s_0$ is given by

$$F_{Y_i^*,Y_i,T_i}\left(\gamma_1, y_i, t_i\right) = \Phi_3\left(\gamma_1 - \mathbf{x}_{zi}'\beta, \frac{y_i - \mathbf{x}_{yi}'\beta_1}{\sigma}, \Phi^{-1}\left(F_T\left(t_i\right)\right); R\right)$$

Therefore (vide appendix)

$$\begin{aligned} f\left(Z_i = s_0, y_i, t_i\right) &= \frac{1}{\sigma} f_{T_i}\left(t_i\right) \Phi\left(\frac{\gamma_1 - \mathbf{x}_{z_i}'\beta - R_{12}R_{22}^{-1}\left(\frac{y_i - \mathbf{x}_{y_i}'\beta_1}{\sigma}, \Phi^{-1}\left(F_{T_i}\left(t_i\right)\right)\right)}{\sqrt{1 - R_{12}R_{22}^{-1}R_{21}}}\right) \\ &\quad \times \phi\left(\frac{\frac{y_i - \mathbf{x}_{y_i}'\beta_1}{\sigma} - \rho_{yt}\Phi^{-1}\left(F_{T_i}\left(t_i\right)\right)}{\sqrt{1 - \rho_{yt}^2}}\right). \end{aligned}$$

The other joint distribution except $F_{Y_i,T_i}\left(y_i, t_i\right)$ are obtained in the same manner. For $Z_i = s_c$ (vide appendix)

$$F_{Y_i,T_i}\left(y_i, t_i\right) = \Phi_2\left(\left[\frac{y_i - \mathbf{x}_{y_i}'\beta_1}{\sigma}, \Phi^{-1}\left(F_{T_i}\left(t_i\right)\right); R_{22}\right]\right)$$

and

$$\begin{aligned} f\left(y_i, t_i\right) &= \frac{1}{\sigma} f_{T_i}\left(t_i\right) \Phi\left(\frac{y_i - \mathbf{x}_{y_i}'\beta_1}{\sigma} | \Phi^{-1}\left(F_{T_i}\left(t_i\right)\right)\right) \\ &= \frac{1}{\sigma} f_{T_i}\left(t_i\right) \phi\left(\frac{\frac{y_i - \mathbf{x}_{y_i}'\beta_1}{\sigma} - \rho_{yt}\Phi^{-1}\left(F_{T_i}\left(t_i\right)\right)}{\sqrt{1 - \rho_{yt}^2}}\right). \end{aligned}$$

In the same manner, the joint density function of Z_i, Y_i, and T_i which is $f_{Z_i,Y_i,T_i}\left(z_i, y_i, t_i\right) = \frac{\partial^2 P\left(Z_i = z_i, Y_i \le y_i, T_i \le t_i\right)}{\partial y_i \partial t_i}$, is given by

$$f_{Z_i,Y_i,T_i}\left(z, y, t\right) = \begin{cases} \frac{1}{\sigma} f_T\left(t\right) \Phi^*\left(\gamma_1\right) \times \phi\left(\frac{\frac{y - x_y'\beta_1}{\sigma} - \rho_{yt}\Phi^{-1}\left(F_T\left(t\right)\right)}{\sqrt{1 - \rho_{yt}^2}}\right), & Z_i = s_0 \\ \frac{1}{\sigma} f_T\left(t\right) \left[\Phi^*\left(\gamma_2\right) - \Phi^*\left(\gamma_1\right)\right] \times \phi\left(\frac{\frac{y - x_y'\beta_1}{\sigma} - \rho_{yt}\Phi^{-1}\left(F_T\left(t\right)\right)}{\sqrt{1 - \rho_{yt}^2}}\right), & Z_i = s_1 \\ \vdots & \vdots \\ \frac{1}{\sigma} f_T\left(t\right) \left[\Phi^*\left(\gamma_{k+1}\right) - \Phi^*\left(\gamma_k\right)\right] \times \phi\left(\frac{\frac{y - x_y'\beta_1}{\sigma} - \rho_{yt}\Phi^{-1}\left(F_T\left(t\right)\right)}{\sqrt{1 - \rho_{yt}^2}}\right), & Z_i = s_k \\ \vdots & \vdots \\ \frac{1}{\sigma} f_T\left(t\right) \left[1 - \Phi^*\left(\gamma_C\right)\right] \times \phi\left(\frac{\frac{y - x_y'\beta_1}{\sigma} - \rho_{yt}\Phi^{-1}\left(F_T\left(t\right)\right)}{\sqrt{1 - \rho_{yt}^2}}\right), & Z_i = s_C \end{cases} \quad (3)$$

where $\Phi^*(\gamma_k) = \Phi\left(\frac{\gamma_k - \mathbf{x}_z'\beta - R_{12}R_{22}^{-1}\left(\frac{y-\mathbf{x}_y'\beta_1}{\sigma}, \Phi^{-1}(F_T(t))\right)}{\sqrt{1-R_{12}R_{22}^{-1}R_{21}}}\right)$ and ϕ is the standard normal density. By Eq. (3), the likelihood function considering sampling weights is given by

$$\begin{aligned} lnL = l\left(\Theta|z,y,t,w\right) &= \sum_{i=1}^{n} w_i lnf_{Z_i,Y_i,T_i}(z_i,y_i,t_i) \\ &= \sum_i I_{\{s_k\}}(z_i) w_i ln\left[\frac{1}{\sigma} f_{T_i}(t_i)\Phi\left(\frac{\gamma_{k+1} - \mathbf{x}_{z_i}'\beta - R_{12}R_{22}^{-1}\left(\frac{y_i-\mathbf{x}_{y_i}'\beta_1}{\sigma}, \Phi^{-1}\left(F_{T_i}(t_i)\right)\right)}{\sqrt{1-R_{12}R_{22}^{-1}R_{21}}}\right)\right. \\ &\quad \left. - \Phi\left(\frac{\gamma_k - \mathbf{x}_{z_i}'\beta - R_{12}R_{12}^{-1}\left(\frac{y_i-\mathbf{x}_{y_i}'\beta_1}{\sigma}, \Phi^{-1}\left(F_{T_i}(t_i)\right)\right)}{\sqrt{1-R_{12}R_{22}^{-1}R_{21}}}\right)\right] \\ &\quad + \sum_{i=1}^{n} w_i log\phi\left(\frac{\frac{y_i-\mathbf{x}_{y_i}'\beta_1}{\sigma} - \rho_{yt}\left(\Phi^{-1}\left(F_{T_i}(t_i)\right)\right)}{\sqrt{1-\rho_{yt}^2}}\right). \end{aligned} \tag{4}$$

where w_i is the sampling weights of the ith individual, and $I_{\{s_k\}}(z_i)$ is an indicator function which is one where $z_i = s_k$ and 0 otherwise for $k = 0, ..., C$. In sampling survey, the data on each member of the sample will be weighed to be a better representor of the population. In order to obtain unbiased parameters estimates, sampling weights should be considered in the likelihood function.

We also obtain the likelihood function for mixed responses where one of the responses, time to event outcome, may have censored values. In this case, T_i denote the observed event time data for the ith subject, which is taken as the minimum of the true event time T_i^* and the censoring time, C_i. Let censuring indicator, $\delta_i = I\left(T_i^* \leq C_i\right)$, be 0 for right-censored and 1 for observed individuals. Therefore, the observed data for time outcome consist of the pairs (T_i, δ_i), $i = 1, 2, ..., n$.

Let Θ be the vector of all unknown parameters of the model. The likelihood function is given by

$$L\left(\Theta|z,y,t,\delta,w\right) = \prod_{i=1}^{n}\left[f_{Z_i,Y_i,T_i}^{\delta_i}\left(z_i,y_i,t_i\right) S_i^{(1-\delta_i)}\left(z_i,y_i,C_i\right)\right]^{w_i}$$

where $S_i\left(z_i,y_i,C_i\right) = P\left(Z_i = z_i, Y_i \leq y_i, T_i > C_i\right)$. The log-likelihood function is given by

$$l\left(\Theta|z,y,t,\delta,w\right) = \sum_{i=1}^{n} w_i\delta_i lnf_{Z_i,Y_i,T_i}\left(z_i,y_i,t_i\right) + \sum_{i=1}^{n} w_i\left(1-\delta_i\right) lnS_i\left(z_i,y_i,C_i\right).$$

This likelihood is different with the likelihood given in Eq. (4), as if consider censoring for some outcomes of the time to event.

4. SIMULATION STUDY

In this section, some simulation studies are conducted to illustrate the performance of the proposed model. The structures, that will be used in this section, are considered to be similar to structures we need for analyzing our real data set. In this simulation study, we simulate data from a correlated mixed outcomes including continuous, time to event, and ordinal responses. Also, the effect of varying values of correlation matrix are considered. In another simulation study, the effect of using Gaussian copula on the data generated by non-Gaussian copulas such as Clayton, Frank, Gumbel, and t-copula are also investigated. In this section, two kinds of dependence (Kendall's τ and Spearman's ρ) are captured by Gaussian copula.

Two simulation studies with sample size $n = 200$ and $n = 1000$ are considered and 500 iterations are performed. For the continuous response, the following identity link function is used:

$$\mu_{i1} = \beta_{11} + \beta_{12}X_{1i} + \beta_{13}X_{2i},$$

where we consider $\beta_1 = (\beta_{11}, \beta_{12}, \beta_{13}) = (2, 1, 3)$. The X_{1i}s are generated from a normal distribution with mean 0 and variance $\sigma^2 = 1$ and X_{2i}s are generated from a Bernoulli distribution with success probability 0.3.

Also, a Weibull model is considered for the time to event outcome with the following linear predictor:

$$log\eta_i = \beta_{21} + \beta_{22}X_{1i}.$$

where $\beta_2 = (\beta_{21}, \beta_{22}) = (2, -2)$. Here the scale parameter is considered as $r = 2$. We also consider an ordinal variable with three categories with cut points $\gamma_1 = 2$ and $\gamma_2 = 3$ as defined in the following equation:

$$Z_i = \begin{cases} 1 & Y_i^* < 2, \\ 2 & 2 \leq Y_i^* < 3, \\ 3 & Y_i^* \geq 3. \end{cases}$$

where the continuous latent variable model for generating ordinal response has the following link function

$$\mu_i = \beta X_{1i}.$$

where we consider $\beta = 3$. We linked these three models using a multivariate distribution constructed from a Gaussian copula with normal marginal distributions for the continuous and the latent variable of the ordinal variable and with a Weibull marginal for the time event outcome using the following positive define matrix:

$$R = \begin{bmatrix} 1 & \rho_{zy} & \rho_{zt} \\ \rho_{zy} & 1 & \rho_{yt} \\ \rho_{zt} & \rho_{yt} & 1 \end{bmatrix},$$

with components $\rho_{zy} = 0.4$, $\rho_{zt} = 0.2$ and $\rho_{yt} = 0.6$.

As mentioned before, copulas provide a natural way to study and measure dependency between random variables. In order to analyze dependence of multivariate distributions, two important measures of dependence known as Kendall's τ and Spearman's o need to be explained. For a multivariate distribution, one can consider the set of bivariate Kendall's τ and Spearman's ρ dependency coefficients. If C is a bivariate copula, Kendall's τ and Spearman's ρ can be represented in semiclosed form, respectively, as

$$\tau = 4\int_{[0,1]^2} CdC - 1 = 1 - 4\int_{[0,1]^2} \frac{\partial C}{\partial u}(u,v)\frac{\partial C}{\partial v}(u,v)\,dudv,$$

and

$$\begin{aligned} \rho_S &= 12\int_{[0,1]^2} uvdC(u,v) - 3 = 12\int_{[0,1]^2} C(u,v)dudv - 3 \\ &= 3 - 12\int_{[0,1]^2} u\frac{\partial C}{\partial u}(u,v)dudv, \end{aligned}$$

For the bivariate Gaussian distribution $\Phi_2(.;\rho)$, the Kendall's τ and Spearman's ρ have closed forms and are defined as

$$\tau = \frac{2}{\pi}arcsin(\rho)$$

$$\rho_S = \frac{6}{\pi}arcsin\left(\frac{\rho}{2}\right).$$

In this section, for three random variables Z, Y, and T, the pattern of correlations of pairs (Z, Y), (Z, T), and (Y, T) should be defined. In a trivariate distribution, three possible bivariate dependency structures for Kendall's τ and Spearman's ϱ values of simulated data should be reported. In this simulation study and the application part, these dependence structures will be presented.

The results of these simulation studies are presented in Table 2. This table contains estimated values of parameters, S.Es.), relative biases (Rel. Bias), and mean square errors (MSEs). These criteria are define as

$$Rel.Bias(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{\theta}_i}{\theta} - 1\right),$$

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\theta}_i - \theta\right)^2,$$

Table 2 | Results of simulation study, mean (Est.), S.E., Rel. Bias, and MSE of parameter estimate under proposed model with considering covariates using 500 iterations of sample size of 200 and 1000.

		$N = 200$				$N = 1000$			
Par	**Real Value**	**Est.**	**S.E.**	**MSE**	**Rel. Bias**	**Est.**	**S.E.**	**MSE**	**Rel. Bias**
β	3	3.171	0.464	0.244	0.057	3.021	0.1725	0.030	0.007
β_{11}	2	1.993	0.074	0.005	−0.003	2.000	0.036	0.001	0.000
β_{12}	1	1.004	0.067	0.004	0.004	0.998	0.030	0.001	−0.001
β_{13}	3	2.997	0.121	0.015	−0.001	2.998	0.055	0.003	−0.000
β_{21}	2	2.029	0.117	0.015	0.015	2.006	0.051	0.002	0.003
β_{22}	−2	−2.023	0.131	0.017	0.015	−2.000	0.058	0.003	0.000
r	2	2.024	0.113	0.013	0.012	2.003	0.050	0.002	0.002
σ	1	0.994	0.050	0.002	−0.005	0.998	0.021	0.000	−0.002
ρ_{zy}	0.4	0.412	0.126	0.016	0.031	0.403	0.055	0.003	0.008
ρ_{zt}	0.2	0.202	0.144	0.021	0.008	0.204	0.058	0.003	0.021
ρ_{yt}	0.6	0.600	0.048	0.002	0.000	0.600	0.019	0.000	0.001
γ_1	2	2.109	0.336	0.125	0.054	2.018	0.126	0.016	0.009
$\gamma_2 - \gamma_1$	1	1.049	0.233	0.057	0.049	1.003	0.090	0.008	0.003

Est., estimate; MSE, mean of square error; Rel. Bias, relative bias; S.E., standard error.

Table 3 | Results of simulation study, mean (Est.), S.E., Rel. Bias, and MSE of parameter estimate under varying values of correlation matrix with components $\rho = (\rho zy, \rho zt, \rho yt)$, using 500 iterations of sample size of 200.

		$\rho = (0.1, 0.2, 0.1)$				$\rho = (0.4, 0.5, 0.5)$				$\rho = (0.7, 0.8, 0.6)$			
Par	**Real Value**	**Est.**	**S.E.**	**MSE**	**Rel. Bias**	**Est.**	**S.E.**	**MSE**	**Rel. Bias**	**Est.**	**S.E.**	**MSE**	**Rel. Bias**
β	3	3.126	0.440	0.209	0.042	3.101	0.424	0.189	0.033	3.142	0.359	0.149	0.048
β_{11}	2	2.000	0.088	0.008	0.000	2.000	0.079	0.006	0.000	1.993	0.085	0.007	−0.003
β_{12}	1	0.999	0.073	0.005	−0.000	0.999	0.069	0.005	−0.001	1.002	0.071	0.005	0.002
β_{13}	3	2.998	0.155	0.024	−0.001	2.997	0.129	0.0166	−0.001	3.006	0.129	0.016	0.002
β_{21}	2	2.026	0.118	0.014	0.013	2.014	0.120	0.014	0.007	2.029	0.113	0.013	0.014
β_{22}	−2	−2.016	0.129	0.017	0.008	−2.007	0.132	0.017	0.004	−2.018	0.126	0.016	0.009
r	2	2.021	0.111	0.013	0.010	2.014	0.114	0.0134	0.007	2.020	0.109	0.012	0.010
σ	1	0.991	0.051	0.003	−0.009	0.991	0.0511	0.003	−0.009	0.989	0.048	0.002	−0.011
ρ_{zy}	ρ_{zy}	0.098	0.147	0.022	−0.021	0.412	0.121	0.015	0.030	0.709	0.074	0.005	0.013
ρ_{zt}	ρ_{zt}	0.195	0.147	0.022	−0.023	0.512	0.112	0.013	0.024	0.809	0.061	0.004	0.011
ρ_{yt}	ρ_{yt}	0.092	0.071	0.005	−0.079	0.501	0.055	0.003	0.003	0.596	0.047	0.0022	−0.005
γ_1	2	2.082	0.324	0.111	0.041	2.072	0.311	0.101	0.036	2.094	0.276	0.085	0.047
$\gamma_2 - \gamma_1$	1	1.041	0.227	0.053	0.041	1.026	0.226	0.052	0.026	1.045	0.2055	0.0445	0.045

Est., estimate; MSE, mean of square error; Rel. Bias, relative bias; S.E., standard error.

$$S.E.(\theta) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\widehat{\theta}_i - \overline{\widehat{\theta}}\right)^2}.$$

where θ is the parameter, $\widehat{\theta}_i$ is its estimate, and $\overline{\widehat{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i$. The results show that parameters are well estimated and by increasing sample size, the values of MSEs are decreased. This suggests that the proposed approach gives consistent estimates.

In order to investigate the effect of various model misspecifications (under varying values of correlation matrix, or under a non-Gaussian copula) on the results, some other simulation studies are performed. We use other positive define matrices R with components $\rho = (0.1, 0.2, 0.1)$, $\rho = (0.4, 0.5, 0.5)$, and $\rho = (0.7, 0.8, 0.6)$. The results of these simulation studies with using 500 iterations of sample size 200 are presented in Table 3.

In Table 4, Kendall's τ and ρ_S values of sample pairs (Z, Y), (Z, T), and (Y, T), under different values of correlation matrix with components $\rho = \left(\rho_{zy}, \rho_{zt}, \rho_{yt}\right)$, based on the result of simulation of Table 3, are given. Comparison of Kendall's τ and ρ_S values in this table show that ρ_S is larger than Kendall's τ. Also by increasing the values of components of the correlation matrix $\rho = \left(\rho_{zy}, \rho_{zt}, \rho_{yt}\right)$, τ and o_S correlations increase.

The results of simulation studies under a non-Gaussian copula with using 500 iterations of sample size 200 are also presented in Table 5. Generating data by different copulas [31] such as Clayton, Frank, Gumbel, and t-copula and analyzing them by Gaussian copula show that the generated data by Clayton and T are less sensitive than those of Gumbel and Frank copulas. However, correlation parameters are estimated with some biases using Gaussian copula.

In Table 6, Kendall's τ and ρ_S values of sample pairs (Z, Y), (Z, T), and (Y, T), under generating data by different copulas and analyzing them by Gaussian copula, based on the results of Table 5, are given. Comparison of Kendall's τ and ρ_S in this table, also show that ρ_S is larger than Kendall's τ. Also, for different copulas, increasing the components of the correlation matrix $\rho = (\rho_{zy}, \rho_{zt}, \rho_{yt})$, leads toincrease of the τ and ρ_S values of sample pairs (Z, Y), (Z, T), and (Y, T).

Table 4 Kendall's tau (τ) and Spearman's ρ (ρS) values of sample pairs (Z, Y), (Z, T) , and (Y, T) calculated under different values of correlation matrix with components ρ = (ρzy, ρzt, ρyt) , using 500 iterations of sample size of 200.

	Real value								
	$\rho = (0.1, 0.2, 0.1)$			$\rho = (0.4, 0.5, 0.5)$			$\rho = (0.7, 0.8, 0.6)$		
Par	Est. (S.E.)	τ	ρ_S	Est.(S.E.)	τ	ρ_S	Est.(S.E.)	τ	ρ_S
ρ_{zy}	0.098 (0.147)	0.0624	0.0936	0.412 (0.121)	0.2703	0.3962	0.709 (0.074)	0.5017	0.6921
ρ_{zt}	0.195 (0.147)	0.1249	0.1865	0.512 (0.112)	0.3422	0.4944	0.809 (0.061)	0.5999	0.7953
ρ_{yt}	0.092 (0.071)	0.0586	0.0878	0.501 (0.055)	0.3340	0.4835	0.596 (0.047)	0.4064	0.5779

Est., estimate; S.E., standard error.

Table 5 Results of simulation study, mean (Est.), S.E., Rel. Bias, and MSE of parameter estimate under different copula with considering covariates using 500 iterations of sample size of 200.

		T				Clayton copula				Frank copula				Gumbel copula			
Par	Real value	Est.	S.E.	MSE	Rel. Bias	Est.	S.E.	MSE	Rel. Bias	Est.	S.E.	MSE	Rel. Bias	Est.	S.E.	MSE	Rel. Bias
β	3	3.316	0.735	0.639	0.105	3.251	0.651	0.486	0.083	3.360	0.834	0.824	0.120	3.271	0.625	0.464	0.090
β_{11}	2	2.003	0.115	0.013	0.001	1.989	0.115	0.013	−0.005	1.996	0.120	0.014	−0.001	2.000	0.115	0.013	0.000
β_{12}	1	0.994	0.100	0.010	−0.006	1.002	0.100	0.010	0.002	0.997	0.110	0.012	−0.003	0.999	0.092	0.008	−0.000
β_{13}	3	2.989	0.213	0.045	−0.003	3.004	0.213	0.045	0.001	2.991	0.217	0.047	−0.002	3.008	0.167	0.028	0.002
β_{21}	2	2.045	0.170	0.031	0.023	2.024	0.168	0.028	0.012	2.049	0.166	0.030	0.024	2.070	0.162	0.031	0.035
β_{22}	−2	−2.033	0.196	0.039	0.016	−2.021	0.200	0.040	0.010	−2.054	0.201	0.043	0.027	−2.062	0.186	0.038	0.031
r	2	2.038	0.171	0.031	0.019	2.017	0.163	0.027	0.008	2.048	0.171	0.031	0.024	2.066	0.164	0.031	0.033
σ	1	0.978	0.071	0.005	−0.021	0.979	0.070	0.005	−0.020	0.985	0.067	0.004	−0.015	0.980	0.071	0.005	−0.019
ρ_{zy}	-	0.274	0.238	0.073	−0.315	0.296	0.201	0.051	−0.259	0.083	0.229	0.152	−0.791	0.725	0.013	0.123	0.812
ρ_{zt}	-	0.088	0.248	0.074	−0.558	0.295	0.203	0.050	0.475	0.064	0.228	0.071	−0.679	0.722	0.125	0.288	2.609
ρ_{yt}	-	0.299	0.109	0.102	−0.500	0.318	0.102	0.090	−0.469	0.077	0.097	0.282	−0.870	0.693	0.058	0.012	0.155
γ_1	2	2.199	0.538	0.328	0.099	2.144	0.479	0.249	0.072	2.228	0.613	0.427	0.114	2.169	0.441	0.222	0.084
$\gamma_2 - \gamma_1$	1	1.082	0.342	0.124	0.082	1.074	0.320	0.107	0.074	1.099	0.3714	0.147	0.099	1.089	0.320	0.110	0.089

Est., estimate; MSE, mean of square error; Rel. Bias, relative bias; S.E., standard error.

Table 6 Kendall's tau (τ) and Spearman's ρ (ρS) values of sample pairs (Z, Y), (Z, T) , and (Y, T) calculated under different copula with considering covariates using 500 iterations of sample size of 200.

	T			Clayton copula			Frank copula			Gumbel copula		
Par	Est. (S.E.)	τ	ρ_S	Est.(S.E.)	τ	ρ_S	Est.(S.E.)	τ	ρ_S	Est. (S.E.)	τ	ρ_S
ρ_{zy}	0.274 (0.238)	0.176	0.262	0.296 (0.201)	0.191	0.284	0.083 (0.229)	0.053	0.079	0.725 (0.013)	0.516	0.708
ρ_{zt}	0.195 (0.088)	0.056	0.084	0.512 (0.295)	0.190	0.283	0.064 (0.228)	0.041	0.061	0.722 (0.125)	0.513	0.705
ρ_{yt}	0.092 (0.299)	0.193	0.286	0.501 (0.318)	0.206	0.305	0.077 (0.097)	0.049	0.073	0.693 (0.058)	0.487	0.675

Est., estimate; S.E., standard error.

5. APPLICATION TO THE HOUSEHOLD'S UHS DATA

In this section, we use the household's UHS data implemented by SCI in 2015 to illustrate our methodology. Three interested variables, length of stay in hospital, hospital costs, and results of hospitalization are considered as correlated mixed variables. The data involve N =2486 patients of different ages who are hospitalized. The interest lies in simultaneously linking three outcomes, namely, logarithm of hospital costs, Y_i, (continuous), length of stay in hospital, T_i (continuous), and results of hospitalization, Z_i (ordinal), to the common covariates such as demographic characteristics of patients, that is, gender, age, economic activity status, and other characteristics such as the type of hospital and the type of received health service. Due to small sample size in some levels of the variable services, three levels of specification, treatment and surgery services, for this variable, are merged.

We assume $Y_i \sim N\left(x'_{i1}\beta_1, \sigma\right)$ and $T_i \sim Weibull\,(\eta_i, r)$. We assume $Y_i^* \sim N\left(x'_{zi}\beta, 1\right)$ where Y^* is the unobserved continuous latent variable underlying Z_i (our discrete ordinal response) and Y^* shows the propensity of patient towards important of her/his health. The more the Y^*

is the more is the value of Z. Now consider the following marginal linear models (the initial values of the parameter estimates for using the joint model are chosen by the results of analysis of separate models using available functions in R software):

$$\begin{aligned}
\mu_i(\mathbf{x}_{zi},\beta) &= \beta_1 Age + \beta_2 Gender + \beta_3 Activity_1 + \beta_4 Activity_2 + \beta_5 Service_1 \\
&\quad + \beta_6 Service_2 + \beta_7 Service_3 \\
\mu_{i1}(\mathbf{x}_{yi},\beta_1) &= \beta_{11} + \beta_{12} Age + \beta_{13} TH_1 + \beta_{14} TH_2 + \beta_{15} Area \\
&\quad + \beta_{16} Activity_1 + \beta_{17} Activity_2 + \beta_{18} Service_1 + \beta_{19} Service_2 + \beta_{1,10} Service_3 \\
log(\eta_i) &= \beta_{21} + \beta_{22} Gender + \beta_{23} TH_1 + \beta_{24} TH_2 + \beta_{25} Literacy_1 + \beta_{26} Literacy_2 \\
&\quad + \beta_{27} Literacy_1 * Gender + \beta_{28} Literacy_2 * Gender + \beta_{29} Service_1 \\
&\quad + \beta_{2,10} Service_2 + \beta_{2,11} Service_3
\end{aligned}$$

where for nominal covariates, the usual dummy variables are used, TH is the type of hospital, $\mathbf{x}_{zi}$= (Age, Gender, Economic Activity status, Service), $\mathbf{x}_{yi}$= (Age, TH, Area of Residence, Economic Activity, Service), and $\mathbf{x}_{ti}$= (Gender, TH, Literacy Status, Service). Two models are used and based on the following criteria (Akaike Information Criterion [AIC], Bayesian information criterion [BIC], and Hannan-Quinn information criterion [HQC]):

$$\begin{aligned}
AIC &= -2logl + 2p \\
BIC &= -2logl + plogn \\
HQC &= -2logl + 2ln(lnn)
\end{aligned}$$

are compared. These models are the overall model in Eq. (4), Model I (joint model), and a model that does not consider the correlation, Model II (separate model). For obtaining the maximum likelihood [ML] estimates, one needs an unconstrained and a box-constrained optimization method [32] using PORT routines written by David Gay at Bell Labs (imported to R by Douglas Bates). We use an algorithm using PORT routine by R software. Also, standard deviation of parameter estimates are computed by the inverse of the observed Hessian matrix. As results show, Model I has a better fit than Model II based on values of AIC, BIC, and HQC.

The results of using Model I show that as age increases, the probability of having ineffective admitted increases. Males are more likely to fully or partially recovered than females. Employed people are more likely to fully or partially recovered than inactive and unemployed people. People with service of child birth are more likely to fully or partially recovered than other people. The results of using Model I show that the cost of hospitalization for older people is more than that of younger people. Also, the cost of hospitalization in private and government hospitals are more than those of other hospitals. The costs of hospitalization in private hospitals are more than that of government hospitals. Cost of hospitalization in rural area is less than that of urban area. Also, the cost of hospitalization of employed and unemployed people is more than that of inactive people. This cost for employed people are more than that of unemployed people.

Males have longer duration of hospitalization than that of females. Duration of hospitalization in government hospitals is shorter than that of other hospitals. But, duration of hospitalization in private hospitals is longer than that of other hospitals. Also people who are illiterate or diploma have shorter duration of hospitalization than that of people with higher education. People who received medico, rehabilitation, specification, treatment, or surgery services, have shorter duration of hospitalization than those of people who received child birth services. The results of Table 7 show that three response variables are significantly correlated. Also, the Kendall's τ and ρ_S values of sample pairs (Z, Y), (Z, T), and (Y, T) are, respectively, $\tau = (0.0920, 0.075, 0.1824)$ and $\rho_S = (0.1377, 0.1128, 0.2707)$ which show the positive correlation.

In order to compare the different models, the AIC is computed for joint and separate models in Table 7. Given any two or more estimated models, the model with the lowest value of AIC is the one to be preferred. The results of Table 7 shows that Model I (joint model), has a better fit of the data.

The predictive plot of survival curve of duration of hospitalization in different combination levels of gender and literacy in Fig. 6, obtained by the results of fitted Model I, shows the same pattern as that of Fig. 5 which emphasizes the well performance of the model. In this plot, the values of other covariates are prefixed. We consider gender as a male, type of hospital as a governmental, and services as rehabilitation.

In order to estimate the benefit of increasing literacy, in terms of raising the probability of a full recovery in relatively short time of stay in hospital and low cost of the hospitalizations, the conditional probability of full recovery given short time of stay in hospital and low cost of the hospitalizations should be determined for different levels of literacy. By last line of Eq. (8) and $f(y_i, t_i)$, this probability is

$$\begin{aligned}
f(z_i = s_C | y_i, t_i) &= \frac{f(z_i = s_C, y_i, t_i)}{f(y_i, t_i)} \\
&= 1 - \Phi\left(\frac{\gamma_C - \mathbf{x}'_{z_i}\beta - R_{12}R_{22}^{-1}\left(\frac{y_i - \mathbf{x}'_{y_i}\beta_1}{\sigma}, \Phi^{-1}\left(F_{T_i}(t_i)\right)\right)}{\sqrt{1 - R_{12}R_{22}^{-1}R_{21}}} \right).
\end{aligned}$$

The probability of a full recovery in different combination levels of length of stay in hospital (minimum, mean and maximum) and cost of the hospitalizations (minimum, mean, and maximum) in Fig. 7, obtained by the results of fitted Model I, shows that by increasing the length of stay in hospital and cost of the hospitalizations, the probability of a full recovery will be increased. Analyzing the residuals for the continuous variable of cost shows that the response 42 is an outlier, but removing this individual and analyzing the remaining data does not affect our results.

Table 7 | Parameter Est. and S.Es. obtained by Gaussian copula to construct a joint distribution for three discrete and continuous responses in household's utilization of health services data (Model I: joint model with some interactions effects, Model II: separate models with some interaction effects).

	Model I		**Model II**	
Par (X)	**Est.**	**S.E.**	**Est.**	**S.E.**
			Z_i**: results of hospitalization**	
β_1 (Age)	0.0076	0.0000	0.0080	0.0000
β_2 (Gender: Male)	−0.0810	0.0019	−0.0862	0.0019
Economic activity status (Baseline: Inactive)	-	-	-	-
β_3 (Employed)	−0.2237	0.0021	−0.2282	0.0021
β_4 (Unemployed)	0.0068	0.0035	0.0242	0.0035
Services (Baseline: Child birth)	-	-	-	-
β_5 (Specification, treatment, surgery)	1.9135	0.0040	1.8956	0.0040
β_6 (Medico)	2.0966	0.0040	2.0788	0.0040
β_7 (Rehabilitation)	2.4878	0.0043	2.4712	0.0043
γ_1	2.0362	0.0045	2.0363	0.0045
$\gamma_2 - \gamma_1$	1.6644	0.0013	1.6668	0.0013
			Y_i**: logarithm of hospital costs**	
β_{11}(Intercept)	12.1225	0.0042	12.0664	0.0042
β_{12}(Age)	0.0045	0.0000	0.0061	0.0000
Type of hospital (Baseline: Others)	-	-	-	-
β_{13}(Governmental)	0.3532	0.0028	0.3607	0.0028
β_{14} (Private)	1.6464	0.0033	1.6464	0.0033
Residence area (Baseline: Urban)	-	-	-	-
β_{15}(Rural)	−0.1873	0.0019	−0.1358	0.0020
Economic activity status (Baseline: Inactive)	-	-	-	-
β_{16}(Employed)	0.2311	0.0021	0.2169	0.0022
β_{17}(Unemployed)	0.1536	0.0040	0.2134	0.0042
Services (Baseline: Child birth)	-	-	-	-
β_{18} (Specification, treatment, surgery)	0.5095	0.0029	0.4138	0.0029
β_{19}(Medico)	0.4180	0.0029	0.3240	0.0029
$\beta_{1,10}$(Rehabilitation)	0.2690	0.0035	0.1769	0.0035
σ	1.4545	0.0006	1.4456	0.0006
			T_i**: length of stay in hospital**	
β_{21}(Intercept)	−1.3045	0.0034	−1.3447	0.0035
Gender (Baseline: Female)	-	-	-	-
β_{22}(Male)	0.4810	0.003	0.5068	0.0034
Type of hospital (Baseline: Others)	-	-	-	-
β_{23}(Governmental)	−0.1034	0.0020	−0.1139	0.0020
β_{24}(Private)	0.3022	0.0023	0.3011	0.0023
Literacy status (Baseline: Higher education)	-	-	-	-
β_{25}(Illiterate)	−0.3311	0.0031	−0.3199	0.0032
β_{26} (Diploma)	−0.0940	0.0026	−0.0905	0.0027
Literacy status*Gender(Baseline: Higher education*Male)	-	-	-	-
β_{27}(Illiterate*Male)	−0.2711	0.0041	−0.4928	0.0020
β_{28}(Diploma*Male)	−0.1002	0.0035	−0.5762	0.0020
Services (Baseline: Child birth)	-	-	-	-
β_{29} (Specification, treatment, surgery)	−0.5158	0.0020	−0.8803	0.0025
$\beta_{2,10}$(Medico)	−0.6033	0.0020	−0.2938	0.0042
$\beta_{2,11}$(Rehabilitation)	−0.8984	0.0024	−0.0758	0.0037
r	1.3052	0.0005	1.3073	0.0005
Correlation Structure				
ρ_{zy}	0.1441	0.0007	-	-
ρ_{zt}	0.1181	0.0008	-	-
ρ_{yt}	0.2826	0.0006	-	-
AIC	24,356,940		24,585,384	
BIC	24,357,144		24,585,570	
HQC	24,356,874		24,585,324	

AIC, Akaike Information Criterion; Est., estimate; S.E., standard error.

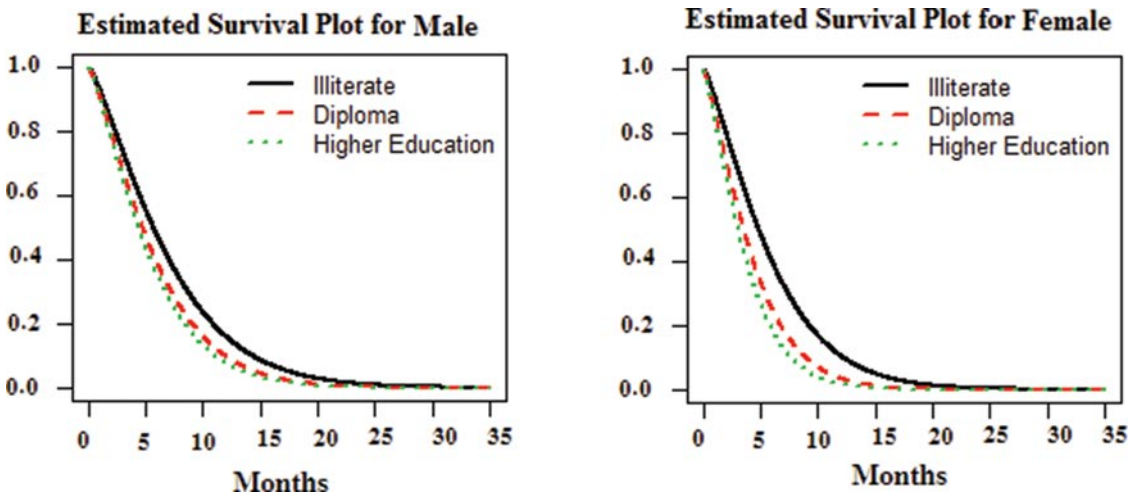

Figure 6 Predictive plot of survival curve of duration of hospitalization in different combination levels of gender and literacy.

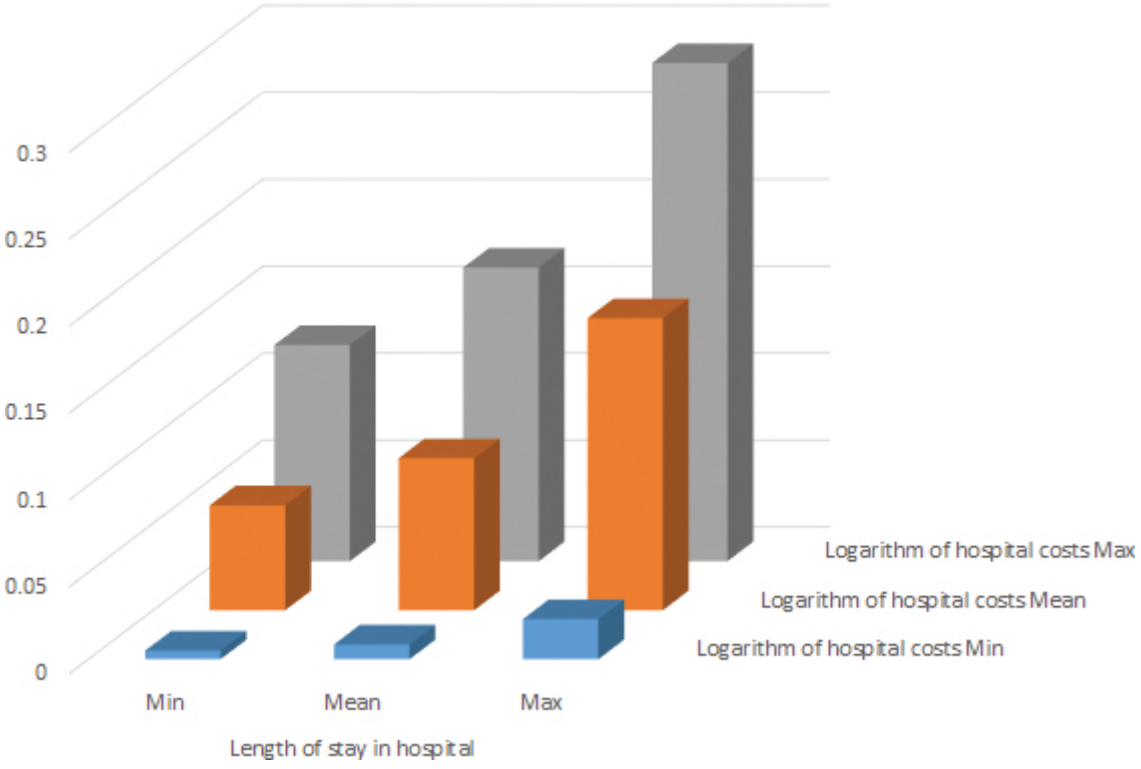

Figure 7 Probability of a full recovery given different combination levels of length of stay in hospital and cost of the hospitalizations.

6. CONCLUSION

In this paper, copula-based regression models for mixed discrete and continuous outcomes, with application in household's UHS data were discussed. Proposed model was also illustrated on some simulated data. It was shown how to construct a joint model for three discrete and continuous responses using Gaussian copula. Likelihood of this model is able to consider the important information of sampling weights. In addition to having continuous and ordinal responses, we may have a recorded nominal response. This paper can be extended to consider cases including a nominal response. For nominal responses, another latent variable should be defined [33].

7. APPENDIX

In this appendix we find the joint trivariate distribution, $f\left(z=s_0,y,t\right)$, and bivariate distribution, $f\left(y,t\right)$, which were used in Section 3. These are, respectively

$$\begin{aligned}
f\left(z=s_0,y,t\right) &= \frac{\partial^2\Phi_3\left(\gamma_1-x_z'\beta,\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right);R\right)}{\partial y\partial t}\\
&= \frac{\partial}{\partial t}\left(\frac{\partial}{\partial y}\int_{-\infty}^{\Phi^{-1}\left(F_T(t)\right)}\int_{-\infty}^{\frac{y-\mathbf{x}_y'\beta_1}{\sigma}}\int_{-\infty}^{\gamma_1-\mathbf{x}_z'\beta}\phi_3(t_1,t_2,t_3)dt_1dt_2dt_3\right)\\
&= \frac{\partial}{\partial t}\left(\int_{-\infty}^{\Phi^{-1}\left(F_T(t)\right)}\int_{-\infty}^{\gamma_1-\mathbf{x}_z'\beta}\frac{1}{\sigma}\phi_3\left(t_1,\frac{y-\mathbf{x}_y'\beta_1}{\sigma},t_3\right)dt_1dt_3\right)\\
&= \frac{1}{\sigma}\int_{-\infty}^{\gamma_1-x_z'\beta}\frac{f_T(t)}{\phi\left(\Phi^{-1}\left(F_T(t)\right)\right)}\phi_3\left(t_1,\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T(t)\right)\right)dt_1\\
&= \frac{1}{\sigma}\frac{f_T\left(t\right)}{\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)}\int_{-\infty}^{\gamma_1-x_z'\beta}\phi\left(t_1|\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T(t)\right)\right)\\
&\quad\times\phi_2\left(\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right)\right)dt_1\\
&= \frac{1}{\sigma}\frac{f_T\left(t\right)}{\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)}\Phi\left(\gamma_1-\mathbf{x}_z'\beta|\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&\quad\times\phi_2\left(\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&= \frac{1}{\sigma}\frac{f_T\left(t\right)}{\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)}\Phi\left(\gamma_1-\mathbf{x}_z'\beta|\frac{y-\mathbf{x}_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&\quad\times\phi\left(\frac{y-\mathbf{x}_y'\beta_1}{\sigma}|\Phi^{-1}\left(F_T\left(t\right)\right)\right)\times\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&= \frac{1}{\sigma}f_T\left(t\right)\Phi\left(\frac{\gamma_1-\mathbf{x}_z'\beta-R_{12}R_{22}^{-1}\left(\frac{y-x_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right)\right)}{\sqrt{1-R_{12}R_{22}^{-1}R_{21}}}\right)\\
&\quad\times\phi\left(\frac{\frac{y-\mathbf{x}_y'\beta_1}{\sigma}-\rho_{yt}\Phi^{-1}\left(F_T\left(t\right)\right)}{\sqrt{1-\rho_{yt}^2}}\right)
\end{aligned}$$

and

$$\begin{aligned}
f\left(y,t\right) &= \frac{\partial^2\Phi_2\left(\frac{y-x_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right);R_{22}\right)}{\partial y\partial t}\\
&= \left(\frac{\partial^2}{\partial y\partial t}\int_{-\infty}^{\Phi^{-1}(F_T(t))}\int_{-\infty}^{\frac{y-x_y'\beta_1}{\sigma}}\Phi_2\left(t_1,t_2\right)dt_1dt_2\right)\\
&= \frac{\partial}{\partial t}\int_{-\infty}^{\Phi^{-1}(F_T(t))}\frac{1}{\sigma}\phi_2\left(\frac{y-x_y'\beta_1}{\sigma},t_2\right)dt_2\\
&= \frac{1}{\sigma}\frac{f_T\left(t\right)}{\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)}\phi_2\left(\frac{y-x_y'\beta_1}{\sigma},\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&= \frac{1}{\sigma}\frac{f_T\left(t\right)}{\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)}\phi\left(\frac{y-\mathbf{x}_y'\beta_1}{\sigma}|\Phi^{-1}\left(F_T\left(t\right)\right)\right)\times\phi\left(\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&= \frac{1}{\sigma}f_T\left(t\right)\phi\left(\frac{y-\mathbf{x}_y'\beta_1}{\sigma}|\Phi^{-1}\left(F_T\left(t\right)\right)\right)\\
&= \frac{1}{\sigma}f_T\left(t\right)\phi\left(\frac{\frac{y-\mathbf{x}_y'\beta_1}{\sigma}-\rho_{yt}\Phi^{-1}\left(F_T\left(t\right)\right)}{\sqrt{1-\rho_{yt}^2}}\right).
\end{aligned}$$

References

1. G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs, Longitudinal Data Analysis, Chapman & Hall/CRC Press, Taylor & Francis Group, Boca Raton, FL, 2008.
2. D.M. Berridge, D.M. Dos Santos, J. Statist. Comput. Simul. 55 (1-2) (1996), 73–86.
3. D.A. Harville, R.W. Mee, Biometrics. 40 (1984), 393-408.
4. G. Verbeke, E. Lesaffre, J. Am. Statist. Assoc. 91 (433) (1996), 217–221.
5. G. Verbeke, G. Molenberghs, Linear Mixed Models in Practice: A SAS Oriented Approach, Springer, New York,1997.
6. J.K. Vermunt, in: J. Hagenaars, A. McCutcheon, Applied Latent Class Analysis, Cambridge University Press, New York, 2002, pp. 383-407.
7. A. Sklar, Fonctions de Repartition ‘a n Dimensions et Leur Marges, Publications de l’Institut de Statistique de L’Universit de Paris 8: 1959; 229-231
8. P. Embrechts, C. Kluppelberg, T. Mikosch, Modelling Extremal Events for Insurance and Finance, Springer, Berlin, 1997.
9. M. Fischer, C. Koeck, S. Schlueter, Quant. Finance. 9 (7) (2009),839–854.
10. M. Ganjali, T. Baghfalaki, J. Biopharm. Stat, 25 (5) (2015), 1077–1099.
11. E. Liebscher, J. Multivariate Anal. 99 (2008), 2234–2250.
12. P.X.-K. Song, M. Li, Y. Yuan, Biometrics. 65 (2009), 60–68.
13. E.C. Brechmannand, U. Schepsmeier, J. Stat. Soft. 52 (3) (2013), 1–27.
14. H. Joe, in: L. Ruschendorf, B. Schweizer, M.D. Taylor, Distributions with Fixed Marginals and Related Topics, Institute of Mathematical Statistics, Hayward, 1996, pp. 120–141.
15. T. Bedford, R.M. Cooke, Ann. Math. Art. Intel. 32 (2001), 245–268.
16. T. Bedford, R.M. Cooke, Ann. Statist. 30 (2002), 1031–1068.
17. D. Kurowicka, R.M. Cooke, Uncertainty Analysis with High Dimensional Dependence Modelling, John Wiley & Sons, Chichester, 2006.
18. A.K. Nikoloulopoulos, D. Karlis, Stat. Med. 27 (2008), 6393–6406.
19. A.R. de Leon, B. Wu, Stat. Med. 30 (2011), 175–185.
20. C.A.J. Klaassen, J.A. Wellner, Bernoulli. 3 (1997), 55–77.
21. G. Masarotto, C. Varin, Electron. J. Stat. 6 (2012), 1517–1549.
22. M.S. Smith, M.A. Khaled, J. Am. Statist. Assoc. 107 (2012), 290–303.
23. I. Žežula, J. Stat. Plan. Infer. 139 (2009), 3942–3946.
24. F. Jiryaie, N. Withanageb, B. Wuc, A.R. de Leon, J. Statist. Comput. Simul. 86 (9) (2016), 1643–1659.
25. J. Stober, H.G. Hong, C. Czado, P. Ghosh, Comp. Statist. Data Anal. 88 (2015), 28–39.
26. A.A. Zilko, D. Kurowicka, Comp. Statist. Data Anal. 103 (2016), 28–55.
27. M. Ram, Appl. Appl. Math. Int. J. 5 (10) (2010), 1483–1492.
28. M. Ram, S.B. Singh, J. Reliab. Stat. Stud. 2 (1) (2009), 91–102.
29. I. Garcia-Subirats, I. Vargas, A.S. Mogolln-Prez, P. De Paepe, M.R.F. da Silva, J.P. Unger, C. Borrell, M.L. Vzquez, Int. J. Equity Health. 13 (10) (2014), 1–15.
30. G.A. Bastos, G.F. Duca, P.C. Hallal, I.S. Santos, Rev. Saude. Publica. 45 (3) (2011), 475–454.
31. J.Yan, J. Stat. Soft. 21(4) (2007), 1–21.
32. P. Fox, The Port Mathematical Subroutine Library, Version 3, AT & T Bell Laboratories, Murray Hill, 1997. http://www.bell-labs.com/project/PORT/
33. L. Fahrmeir, G.T. Tutz, Multivariate Statistical Modelling Based on Generalized Linear Models, Springer-Verlag, Frhwirth-Schnatter S, New York, 1994.