Brazilian Journal of Probability and Statistics
2022, Vol. 36, No. 2, 408–419
https://doi.org/10.1214/22-BJPS532
This research was funded, in whole or in part, by [Slovak Research and Development Agency; Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic, APVV-18-0052 and VEGA 1/0006/19]. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions

A heteroscedasticity diagnostic of a regression analysis with copula dependent random variables

Ayyub Sheikhi[1,a], **Fereshteh Arad**[1,b] **and Radko Mesiar**[2,c]

[1]*Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran,* [a]*sheikhy.a@uk.ac.ir,* [b]*fereshteh_arad@math.uk.ac.ir*
[2]*Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, STU Bratislava, Slovakia,* [c]*radko.mesiar@stuba.sk*

Abstract. One of the most important assumptions in multiple regression analysis is the independence of the explanatory variables, however, this assumption is violated in several situations. In this work, we investigate regression equations when this independence does not hold and the explanatory variables are connected by many of elliptical copulas. We apply the proposed regression equation to study its heteroscedasticity diagnostic and using simulated data we also assess our regression model. A cross-validation procedure is carried out to ensure the unbiasedness of the results. Also, a real data analysis is presented as an application.

1 Introduction and preliminaries

Regression analysis is probably the most popular statistical technique which helps researchers to make a prediction as far as low errors and the linear regression, which is used mostly, is based on multivariate normal assumption of variables. In some cases, the regression equation between explanatory and the output variable is nonlinear. Exponential functions, logarithmic functions, trigonometric functions, power functions, Gaussian functions, and Lorenz curves are examples of such nonlinear functions Seber and Wild (2003), Bates and Watts (2007).

In general, a nonlinear statistical model can be described with the following notation:

$$Y = g(\boldsymbol{X}, \boldsymbol{\theta}) + \varepsilon \tag{1}$$

where Y is the outcome variable, $g(\cdot) : R^p \to R$ is a nonlinear function, $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ are the explanatory variables, $\boldsymbol{\theta}$ denotes the parameters of the model to be estimated, and ε is the error term. The special case of g is the linear form $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ which is called as multiple (linear) regression. Also, g can be a mix of linear and nonlinear functions. There are many techniques that have been proposed by authors which are dealing with nonlinear relations between variables, among them, spline (and B-spline) Marsh and Cormier (2001), radial basis function Ando, Konishi and Imoto (2008), Xu, Krzyżak and Yuille (1994), support vector regression Awad and Khanna (2015), etc. We refer to book of Konishi (2014) for more details about linear and nonlinear models. From another point of view, marginal normal distributions which are connected by a Gaussian copula allow to construct a multivariate normal distribution and hence, their regression equation is linear. Parameter estimations in these regression equations have the same results as the ordinary least squares (OLS) method. In some situations, however, the connection between variables may departure from the Gaussian copula. Analyzing of such models was done in the literature.

Key words and phrases. Gaussian copula, t-copula, regression, heteroscedasticity.
Received February 2021; accepted February 2022.

Crane and van der Hoek (2008) have used some conditional expectation formulae for copulas to carry out some regression analysis. Noh et al. (2013) have presented some inferences of copula-based regressions. Acar, Azimaee and Hoque (2019) have investigated the utility of copula models for model-based predictions. So, exploring the effect of the dependence between the explanatory variables as well as their marginal distributions in a regression equation were our main motivations of this work. For more details in this subject, we refer to Kumar and Shoukri (2007), Hoang, Khandelwal and Ghosh (2019), Bennafla et al. (2016).

Elliptical copulas which are the copula functions corresponding to multivariate elliptical distributions, consist of Gaussian and t-copulas Frahm, Junker and Szimayer (2003).

Denoting Φ the standard normal cumulative distribution and $\boldsymbol{\Phi}_k(., \boldsymbol{R})$ the k-dimensional standard multivariate normal distribution function with correlation matrix $\boldsymbol{R}$, the Gaussian k-dimensional copula, denoted by k-copula, with correlation matrix $\boldsymbol{R}$ is then given by

$$C^{Ga}(u_1, u_2, \ldots, u_k) = \boldsymbol{\Phi}_k\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_k), \boldsymbol{R}\big) \tag{2}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal distribution function and $\boldsymbol{R} \in [-1, 1]^{k\times k}$. In a similar manner, the t-copula is defined as

$$C^t(u_1, u_2, \ldots, u_k) = \boldsymbol{T}_{k,\nu}\big(T_\nu^{-1}(u_1), T_\nu^{-1}(u_2), \ldots, T_\nu^{-1}(u_k), \boldsymbol{R}\big) \tag{3}$$

where $\boldsymbol{T}_{k,\nu}(., \boldsymbol{R})$ is the CDF of the k-dimensional t-distribution with ν degrees of freedom and correlation matrix $\boldsymbol{R}$ and T_ν^{-1} is the inverse of the t-distribution with ν degrees of freedom.

These two copulas play an important role in regression analysis. Standardly, it has been considered that the relation between the explanatory variables and the output variable follow Gaussian copulas as well as the explanatory variables themselves are related via Gaussian copulas, see, for example, Pitt, Chan and Kohn (2006).

Despite the increasing interest in using Gaussian copula for regression analysis, there have been only a few attempts to use t-copula as a tool to carry out a prediction analysis Acar, Azimaee and Hoque (2019), while there exist some cases that these relations can be formulated using a t-copula. In this work, we carry out some regression models when the relation between the explanatory variables follows an elliptical copula. In particular, we first consider a Gaussian copula, then by considering a t-copula, we explore some non-linear equation between the output and input variables. Assuming these two special cases of the elliptical copulas enable us to compare our results with the traditional regression analysis.

Moreover, it is known that, if OLS is performed on a heteroscedastic data set, yielding biased standard error estimation, a researcher might fail to reject a null hypothesis at a given significance level, when that null hypothesis was actually uncharacteristic for the actual population. Cysneiros, Paula and Galea (2007) have investigated heteroscedasticity in linear models and Cysneiros, Cordeiro and Cysneiros (2010) have obtained some ML estimations in heteroscedastic symmetric nonlinear models. Wang and Neal (2012) have presented a Gaussian process regression with heteroscedastic residuals. See also, Kersting et al. (2007) and Alqawba, Diawara and Kim (2019) for more information.

As a motivation and a visual schematic example of homoscedastic versus heteroscedastic regression analysis, we may note to the Figure 1 for $n = 2$ predictors. Figure 1(a) shows fitting a wrong homoscedastic regression analysis on an actually heteroscedastic data set, while, as depicted in Figure 1(b), a heteroscedastic approach will insures the best fitting distribution to the data. From a copula point of view, Chang and Joe (2019) have compared the performance of vine copula and linear regression with conditional heteroscedasticity assumptions. One of our aims in this work is checking the heteroscedasticity in a regression analysis.

The paper is comprised of 4 sections. In next section, we set up the basis of regression equations by means of a copula-based conditional expectation. Some numerical results are presented in Section 3 and finally some concluding remarks are suggested in Section 4.

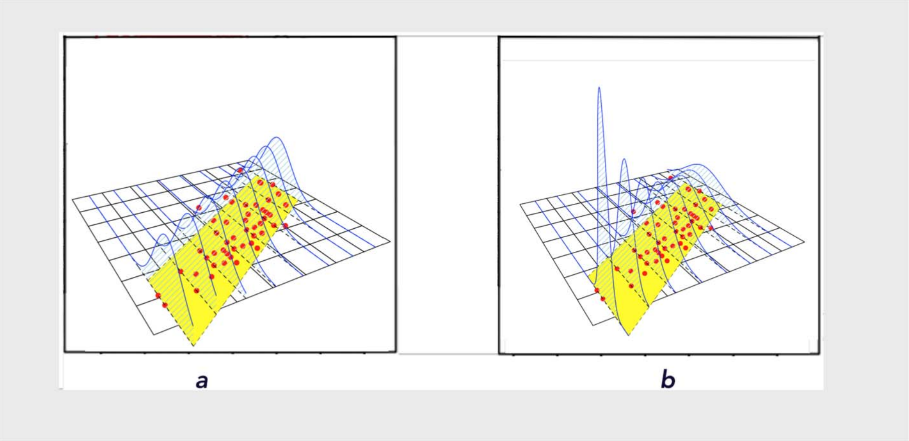

Figure 1 *homoscedastic and heteroscedastic analysis*: (*a*) *homoscedastic*, (*b*) *heteroscedastic*.

2 Copula based conditional expectation

Consider the random vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ is associated with n-copula function $C_{\boldsymbol{X}}(\boldsymbol{u})$ where $\boldsymbol{u} = (u_1, u_2, \dots, u_n)$ and the elements of the random vector $(\boldsymbol{X}, Y)$ are connected by $(n+1)-$ copula $C_{\boldsymbol{X},Y}(\boldsymbol{u}, v)$. The following theorem states a general conditional expectation based on copulas.

Theorem 2.1. *If random variables $X_1, X_2, \dots, X_n, Y$ have marginal distribution functions $G_1(x_1)$, $G_2(x_2)$, $\dots$, $G_n(x_n)$, $F(y)$, respectively, and they are associated with the copula $C_{\boldsymbol{X},Y}(\boldsymbol{u}, v)$, then if exist,*

$$E\big(g(Y)|\boldsymbol{X}=\boldsymbol{x}\big) = \int g(y) \frac{\partial}{\partial y} \frac{D_{123\dots n} C_{\boldsymbol{X},Y}(\boldsymbol{u}, v)}{D_{123\dots n} C_{\boldsymbol{X}}(\boldsymbol{u})} \, dy, \tag{4}$$

where $g(\cdot)$ is a Borel measurable function, $u_i = G_i(x_i)$, $i = 1, 2, \dots, n$ and $v = F(y)$. Also, $D_{123\dots n} C_{\boldsymbol{X},Y}(\boldsymbol{u}, v) = \frac{\partial C_{\boldsymbol{X},Y}(\boldsymbol{u},v)}{\partial u_1 \partial u_2 \dots \partial u_n}$ and $D_{123\dots n} C_{\boldsymbol{X}}(\boldsymbol{u}) = \frac{\partial C_{\boldsymbol{X}}(\boldsymbol{u})}{\partial u_1 \partial u_2 \dots \partial u_n}$ if these derivatives exist, otherwise zero.

Proof. See the Appendix. □

By using the notation $E(g(Y)|\boldsymbol{X}=\boldsymbol{x}) = r_{g(Y)|X_1,\dots,X_n}(y|x_1, \dots, x_n)$, we may obtain the conditional expectation and the conditional variance of Y given $\boldsymbol{X}=\boldsymbol{x}$, respectively, as

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = r_{Y|\boldsymbol{X}}(y|\boldsymbol{x}), \tag{5}$$

$$\operatorname{Var}(Y|\boldsymbol{X}=\boldsymbol{x}) = r_{Y^2|\boldsymbol{X}}(y|\boldsymbol{x}) - r^2_{Y|\boldsymbol{X}}(y|\boldsymbol{x}). \tag{6}$$

A special case of the Theorem 2.1 arises when random variables X_i, $i = 1, 2, \dots, n$ are pairwise independent.

Corollary 2.2. *With the assumptions of Theorem 2.1, if random variables X_1, X_2, $\dots$, X_n are pairwise independent, then if exist,*

$$E\big(g(Y)|\boldsymbol{X}=\boldsymbol{x}\big) = \int g(y) \frac{\partial}{\partial y} D_{123\dots n} C_{\boldsymbol{X},Y}(\boldsymbol{u}, v) \, dy. \tag{7}$$

Proof. The proof is simple noting that under the independence assumption of X_is, $i = 1, 2, \dots, n$, we have

$$D_{123\dots n}C(u_1, u_2, \dots, u_n, 1) = D_{12\dots n}C(u_1, u_2, \dots, u_n)$$
$$= \frac{\partial}{\partial u_1 \partial u_2 \dots \partial u_n} \Pi_{j=1}^{j=n} u_j = 1. \qquad \square$$

We will use the previous results to carry out some regression analysis when the random variables are connected via elliptical copulas. We assume the multiple regression

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon \tag{8}$$

between the explanatory variables $X_1, X_2, \dots, X_n$ and the response variable Y, where β_i, $i = 1, 2, \dots, n$ are the regression coefficients. An immediate application of Theorem 2.1 in this model, yields the results of ordinary least square (OLS) regression.

Corollary 2.3. *If random variables $X_1, X_2, \dots, X_n, Y$ follow standard normal distribution and for $i = 1, 2, \dots, n$, X_i and Y are connected using Gaussian copula with parameter ρ_{iy}, i.e., $C_{X_i,Y}(u, v) = GA(u, v, \rho_{iy})$ and $X_1, X_2, \dots, X_n$ are pairwise independent, then*

$$r_{Y|X_1,\dots,X_n}(y|x_1, \dots, x_n) = \rho_{1y}x_1 + \rho_{2y}x_2 + \cdots + \rho_{ny}x_n. \tag{9}$$

Proof. Since X_j's for $j = 1, \dots, n$ are pairwise independent, we consider the correlation matrix of $\boldsymbol{X}$ and Y as $\boldsymbol{R} = \begin{bmatrix} 1 & \boldsymbol{r}_{yx} \\ \boldsymbol{r}_{yx}^T & \boldsymbol{R}_{xx} \end{bmatrix}$ where $\boldsymbol{r}_{yx} = [\rho_{1y}\ \rho_{2y}\ \dots\ \rho_{ny}]$ the correlation vector between Y and $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ and $\boldsymbol{R}_{xx} = \rho\,\mathrm{diag}(1, 1, \dots, 1)$. By inverse calculation of the correlation matrix, and substituting in the $n+1$-variable Gaussian-copula, we have:

$$C^{Ga}(u_1, u_2, \dots, u_n, v)$$
$$= \int_{-\infty}^{\Phi^{-1}(v)} \oint_{-\infty}^{\Phi^{-1}(u_j)} \frac{1}{\sqrt{(2\pi)^{n+1}(|\boldsymbol{R}|)}}$$
$$\times \exp\biggl(\frac{-1}{2|\boldsymbol{R}|}\bigl((s - (\rho_{1y}t_1 + \cdots + \rho_{ny}t_n))^2 + |\boldsymbol{R}|t_1^2 + \cdots + |\boldsymbol{R}|t_n^2\bigr)\biggr)\, dt_j\, ds,$$

where $\oint$ is an n-integral on t_j, $j = 1, \dots, n$. Then, denoting $\Phi^{-1}(u_1) = b_1, \dots, \Phi^{-1}(u_n) = b_n$ and $\Phi^{-1}(v) = b_0$, and $g(s, t_1, \dots, t_n) = \frac{1}{\sqrt{(2\pi)^{n+1}(|\boldsymbol{R}|)}} \exp(\frac{-1}{2|\boldsymbol{R}|}((s - (\rho_{1y}t_1 + \cdots + \rho_{ny}t_n))^2 + |\boldsymbol{R}|t_1^2 + \cdots + |\boldsymbol{R}|t_n^2))$, by derivation with respect to $u_1, u_2, \dots, u_n$, we obtain

$$D_{123\dots n}C^{Ga}(u_1, u_2, \dots, u_n, v)$$
$$= \frac{\partial}{\partial u_1 \dots \partial u_{n-1}} \frac{\partial b_n}{\partial u_n} \frac{\partial}{\partial b_n} \int_{-\infty}^{b_0} \int_{-\infty}^{b_1} \dots \int_{-\infty}^{b_n} g(s, t_1, \dots, t_n)\, dt_n \dots dt_1\, ds$$
$$= \frac{\partial}{\partial u_1 \dots \partial u_{n-1}} \frac{1}{\phi(b_n)} \int_{-\infty}^{b_0} \int_{-\infty}^{b_1} \dots \left[\frac{\partial}{\partial b_n} \int_{-\infty}^{b_n} g(s, t_1, \dots, t_n)\, dt_n\right] dt_{n-1} \dots dt_1\, ds$$
$$= \frac{\partial}{\partial u_1 \dots \partial u_{n-1}} \int_{-\infty}^{b_0} \int_{-\infty}^{b_1} \dots \int_{-\infty}^{b_{n-1}} g(s, t_1, \dots, b_n)\, dt_{n-1} \dots dt_1\, ds$$
$$= \frac{\partial}{\partial u_1 \dots \partial u_{n-2}} \frac{\partial b_{n-1}}{\partial u_{n-1}} \frac{\partial}{\partial b_{n-1}} \int_{-\infty}^{b_0} \int_{-\infty}^{b_1} \dots \int_{-\infty}^{b_{n-1}} g(s, t_1, \dots, b_n)\, dt_{n-1} \dots dt_1\, ds$$
$$= \frac{\partial}{\partial u_1 \dots \partial u_{n-2}} \frac{1}{\phi(b_{n-1})} \int_{-\infty}^{b_0} \int_{-\infty}^{b_1} \dots \left[\frac{\partial}{\partial b_{n-1}} \int_{-\infty}^{b_{n-1}} g(s, t_1, \dots, b_n)\, dt_{n-1}\right] \dots dt_1\, ds$$

$$= \frac{\partial}{\partial u_1 \dots \partial u_{n-2}} \int_{-\infty}^{b_0} \int_{-\infty}^{b_1} \cdots \int_{-\infty}^{b_{n-2}} g(s, t_1, \dots b_{n-1}, b_n)\, dt_{n-2} \dots dt_1\, ds.$$

After some algebraic computations, we have

$$D_{123\dots n} C^{Ga}(u_1, u_2, \dots, u_n, v) = \Phi\left(\frac{b_0 - (\rho_{1y} b_1 + \cdots + \rho_{ny} b_n)}{\sqrt{|\boldsymbol{R}|}}\right),$$

and substituting values b_0,b_1,...,b_n, it yields

$$D_{123\dots n} C^{Ga}(u_1, \dots, u_n, v) = \Phi\left(\frac{y - (\rho_{1y} x_1 + \cdots + \rho_{ny} x_n)}{\sqrt{|\boldsymbol{R}|}}\right).$$

Hence, equation (7) yields the regression equation as

$$\begin{aligned} r_{Y|X_1,\dots,X_n}(y|x_1, \dots, x_n) &= \int y \frac{\partial}{\partial y} D_{123\dots n} C(u_1, u_2, \dots, u_n, v)\, dy \\ &= \int y \frac{\partial}{\partial y} D_{123\dots n} C^{Ga}(u_1, u_2, \dots, u_n, v)\, dy \\ &= \int y \frac{1}{\sqrt{|\boldsymbol{R}|}} \phi\left(\frac{y - (\rho_{1y} x_1 + \cdots + \rho_{ny} x_n)}{\sqrt{|\boldsymbol{R}|}}\right) dy \\ &= \rho_{1y} x_1 + \rho_{2y} x_2 + \cdots + \rho_{ny} x_n \end{aligned}$$

which is (9). □

As another result of equation (7), we can determine whether the regression equation (9) is heteroscedastic or not. For this, we note that $Y|\boldsymbol{X} \sim N(\boldsymbol{r}_{yx}\boldsymbol{X}, 1 - \boldsymbol{r}_{yx}\boldsymbol{R}_{xx}^{-1}\boldsymbol{r}_{yx}^T)$ So its variance is

$$\operatorname{var}(Y|\boldsymbol{X}) = 1 - \rho_{1y}^2 - \rho_{2y}^2 - \rho_{3y}^2 - \cdots - \rho_{ny}^2,$$

which shows its homoscedastic property. When the independence assumption of explanatory variables is violated, their dependence may be expressed by a copula function. In the sequel, we consider elliptical copulas as a connection function between X_1, X_2 and Y. The detailed solutions of the following examples are presented in the Appendix.

Example 2.4. Assume that the variables Y, X_1, X_2 have marginal standard normal distributions, and they are related by a Gaussian copula, then the regression equation of Y given X_1 and X_2 is

$$r_{Y|X_1,X_2}(y|x_1, x_2) = \frac{\rho_{1y} - \rho_{2y}\rho_{12}}{1 - \rho_{12}^2} x_1 + \frac{\rho_{2y} - \rho_{1y}\rho_{12}}{1 - \rho_{12}^2} x_2. \tag{10}$$

here ρ_{1y}, ρ_{2y} and ρ_{12} is the correlation coefficient of the variable Y with X_1, the correlation coefficient of the variables Y with X_2 and the correlation coefficient of the variable X_1 with X_2, respectively. Also,

$$\operatorname{var}(Y|\boldsymbol{X}) = \frac{1 - \rho_{1y}^2 - \rho_{12}^2 - \rho_{2y}^2 + 2\rho_{1y}\rho_{2y}\rho_{12}}{1 - \rho_{12}^2}.$$

We simply deduce from Corollary 2.3 and above example that in a bivariate regression equation, if two variables X_1 and X_2 are independent and they are connected with Y based on the Gaussian copula, then $\rho_{12} = 0$ and the regression equation of Y given X_1, X_2 will be

$$r_{Y|X_1,X_2}(y|x_1, x_2) = \rho_{1y} x_1 + \rho_{2y} x_2.$$

Another special case of Theorem 2.1 arises when the relation between the explanatory variables is modeled by a t-copula.

Example 2.5. Assume that the random variables Y, X_1, X_2 have marginally standard normal distribution and are associated with t-copula. Then

$$r_{Y|X_1,X_2}(y|x_1,x_2) = \int y \frac{\partial}{\partial y} \frac{D_{12}C^t(u_1,u_2,v)}{D_{12}C^t(u_1,u_2)} dy$$

$$= \int y \frac{\partial}{\partial y} \frac{l}{h} dy,$$

where $h = \frac{1}{t_\nu(m(x_1))t_\nu(m(x_2))2\pi\sqrt{1-\rho_{12}^2}}(1 + \frac{m^2(x_1)-2\rho_{12}m(x_1)m(x_2)+m^2(x_2)}{\nu(1-\rho_{12}^2)})^{\frac{-(\nu+2)}{2}}$ and

$$l = \frac{\Gamma(\frac{\nu+3}{2})|\boldsymbol{R}|^{\frac{-1}{2}}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{3}{2}}t_\nu(m(x_1))t_\nu(m(x_2))} \int_{-\infty}^{m(y)} \Bigg(1 + \frac{1-\rho_{12}^2}{\nu|\boldsymbol{R}|}\Bigg(s - \Bigg(\frac{\rho_{1y}-\rho_{2y}\rho_{12}}{1-\rho_{12}^2}m(x_1)$$

$$+ \frac{\rho_{2y}-\rho_{1y}\rho_{12}}{1-\rho_{12}^2}m(x_2)\Bigg)\Bigg)^2 + \frac{m^2(x_1)+m^2(x_2)-2\rho_{12}m(x_1)m(x_2)}{\nu(1-\rho_{12}^2)}\Bigg)^{\frac{-(\nu+3)}{2}} ds,$$

with $m(u) = T_\nu^{-1}(\Phi(u))$ and $T_\nu^{-1}(\cdot)$ is the inverse of distribution function of t-distribution with ν degrees of freedom.

3 Numerical study

3.1 Simulation study

In this section, we examine our results using a Monte Carlo simulation study.[1] Using R package `Vinecopula` Schepsmeier et al. (2015), we start our simulation with sample size $n = 35$ from a trivariate standard normals (X_1, X_2, Y) in which they are connected via the t-copula with correlation $\rho_{X_1X_2} = 0.5$ and $\rho_{Y,X_i} = 0.6$, $i = 1, 2$. Since we aim to use a 5- fold cross validation, in fact in the kth fold, $k = 1, 2, \ldots, 5$ we have only 28 observations in the train set. We repeated this procedure 1000 times. The mean and standard deviation of Root Mean Square Error (RMSE) in these 5-fold cross-validation was our criterion to compare the accuracy of the copula-based regression against the traditional linear regression. It is well known that RMSE can be obtained as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_j)^2}.$$

Also, the AIC for a regression model can be calculated as follows Bentler (1985)

$$AIC = n * LL + 2 * k,$$

where n is the number of data, LL is the log-likelihood for the model using the natural logarithm (e.g. the log of the MSE), and k is the number of parameters in the model. Moreover, the BIC can be calculated as

$$BIC = n * LL + k * \log(n),$$

where $\log(\cdot)$ is the natural logarithm. Table 1 reveals the means and standard deviations in both training and test sets as well as the AIC and the BIC of test set. As seen from this table, a copula-based regression has two merits. First, the means of RMSEs in the training sets are close, but in test sets, the mean of RMSE of copula regression is less than the classic

[1] We use the R software and all codes are available upon request.

Table 1 *mean and SD of RMSE-train and RMSE-test of t-Copula regression and Linear Regression in simulated dataset*

Estimation	t-Copula regression	Linear Regression
Mean of RMSE-Test	0.70	0.73
SD of RMSE-Test	0.34	0.34
AIC-Test	−101.81	−99.23
BIC-Test	−95.59	−93.01

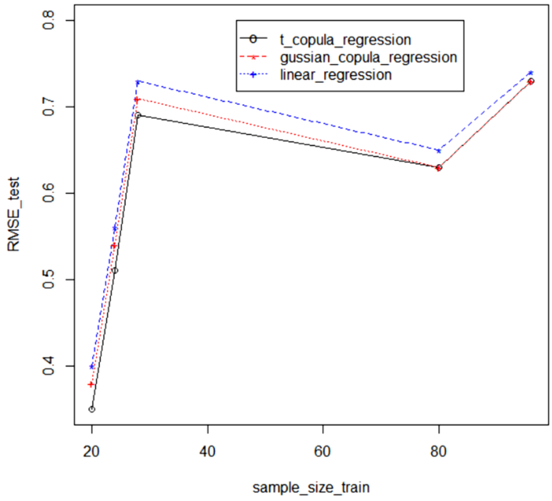

Figure 2 *Performance of ordinary linear regression, t-copula regression and Gaussian copula regression.*

regression which shows the overfitting of the classic regression when the independence of explanatory variables has not met. This is because of the heteroscedasticity of such models, which cannot be captured by a linear regression (see Figure 1(b)). Second, from the standard deviations, we observe that the copula regression gives a robust estimator in favor of the classical regression. Moreover, taking in mind that the t-student distribution (t-copula) converges to the normal distribution (Gaussian copula) as the sample size increases Kole, Koedijk and Verbeek (2007), Figure 2 shows the performance of these three regression analysis: ordinary linear regression, t-copula regression and Gaussian copula regression. Again, we assume that the two random variables X_1, X_2 and Y are connected via a t-copula with correlations coifi-cients $\rho_{X_1X_2} = 0.5$ and $\rho_{Y,X_i} = 0.6$, $i = 1, 2$. We started with $n = 25$ sample size which was divided into 20 observations for training set and 5 observations for testing set. Then, we have increased the sample size to $n = 30$, i.e, 24 sample for training set and 6 samples for testing set. We continued this addition of the sample size until $n = 150$ (120 sample for training set and 30 samples for testing set). As reflected in Figure 2, where the explanatory random variables are connected using a t-copula, the t-copula model is superior to the Gaussian copula model, although both of them significantly perform better than the classical linear regression. Also, as the sample size increases, the RMSEs of the t and Gaussian copula coincide.

3.2 Real data

We used our copula regression to make a prediction in the Parkinson data which was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado Little et al. (2007). The study includes 195 voice recordings from 31 individuals with different numbers of replications of each individual. We were going

Table 2 *Mean and SD of RMSE-test and AIC-Test and BIC-Test of G-Copula regression and Linear Regression in Parkinson dataset*

Estimation	G-Copula regression	Linear Regression
Mean of RMSE-Test	0.01	0.03
SD of RMSE-Test	0.04	0.04
AIC-Test	−373.06	−364.54
BIC-Test	−364.55	−356.04

to predict second nonlinear measures of fundamental frequency variation,“PPE”, using first nonlinear measures of fundamental frequency variation,“DFA”, and a nonlinear dynamical complexity measure,“spread1”. We selected the first two replications of each individuals of these three variables. We first standardized these variables and each indivdual normality was confirmed by the Shapiro-Wilks test of normality. Denoting the standardized values of DFA, spread1 and PPE, respectively, as x_1, x_2, y, we found that the two variables x_1 and x_2 are related with correlation coefficient $\rho_{12} = 0.23$. Also, we observed that y is connected with x_1 with correlation coefficient $\rho_{1y} = 0.24$. Moreover, the connection of y and x_2 is with correlation coefficient $\rho_{2y} = 0.99$. Also, x_1 and x_2 are related with Gaussian copula and y with x_1 and x_2 are connected via Gaussian copula. The effect of x_1 and x_2 on the y using two approaches linear regression and copula regression was carried out applying a 5-fold cross validation. Using a linear regression we obtained regression equation

$$y = 0.005 + 0.01x_1 + 0.98x_2,$$

while using the Gaussian copula regression, we obtained

$$y = 0.01x_1 + 0.99x_2.$$

Table 2 reveals the means and standard deviation of the RMSEs of these two models in the test sets as well as their AIC and BIC. These indices were almost the same for the train sets, except a minor difference in favor of G-Copula regression, but again, their significance differences in the test set claim the superiority of the copula-based regression models.

4 Conclusion

The present study was designed to determine the effect of structural dependences on the regression analysis. We studied some copula-based relations between variables, especially, the Gaussian and t-copulas. The results of this study indicate that considering structural dependences will improve the performance of the regression analysis. In addition, these findings may help us to understand that we are really encountered with a heteroscedastic problem when the explanatory variables are connected using a t-copula. Future studies on the current topic are therefore recommended. Considering some Archimedean copulas instead of these elliptical copulas would be of interest. Also, it might be possible to extend these results with more explanatory random variables and set up a copula-based path analysis. Moreover, although we did a diagnostic research to identify the homoscedasticity/heteroscedasticity of a regression analysis, one may to investigate an idea to handle a heteroscedastic regression analysis using their copula relations. Our next ongoing aim is developing another heteroscedastic analysis, in which, the explanatory random variables are observed with some measurement errors and random noises, see Kim, Li and Spiegelman (2016), Mesiar, Sheikhi and Komorníková (2019), Sheikhi and Mesiar (2020), for instance.

Appendix

Proof of Theorem 2.1

According to the characteristics of the copulas, we have

$$E\big(g(Y)|\boldsymbol{X}=\boldsymbol{x}\big)=\int g(y)\frac{\partial}{\partial y}F_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\,dy$$
$$=\int g(y)\frac{\partial}{\partial y}\frac{\frac{\partial C_{\boldsymbol{X},Y}(\boldsymbol{u},v)}{\partial u_1\partial u_2\ldots\partial u_n}}{\frac{\partial C_{\boldsymbol{X},Y}(\boldsymbol{u},1)}{\partial u_1\partial u_2\ldots\partial u_n}}\,dy$$
$$=\int g(y)\frac{\partial}{\partial y}\frac{D_{123\ldots n}C_{\boldsymbol{X},Y}(\boldsymbol{u},v)}{D_{123\ldots n}C_{\boldsymbol{X}}(\boldsymbol{u})}\,dy,$$

which is (4).

Detailed computation of Example 2.4. We have the 3-Gaussian-copula function as

$$C^{Ga}(u_1,u_2,v)=\int_{-\infty}^{\Phi^{-1}(v)}\int_{-\infty}^{\Phi^{-1}(u_2)}\int_{-\infty}^{\Phi^{-1}(u_1)}\frac{1}{\sqrt{(2\pi)^3|\boldsymbol{R}|}}$$
$$\times\exp\bigg(\frac{-1}{2}(s,t_1,t_2)\boldsymbol{R}^{-1}(s,t_1,t_2)^T\bigg)\,dt_1\,dt_2\,ds,$$

where $v=F(y)$, $u_1=G_1(x_1)$ and $u_2=G_2(x_2)$. Consider the correlation matrix and its partitions are

$$\boldsymbol{R}=\begin{bmatrix}1 & \boldsymbol{r}_{yx}\\ \boldsymbol{r}_{yx}^T & \boldsymbol{R}_{xx}\end{bmatrix},\qquad \mathbf{r}_{yx}=\begin{bmatrix}\rho_{1y} & \rho_{2y}\end{bmatrix},\qquad \boldsymbol{R}_{xx}=\begin{bmatrix}1 & \rho_{12}\\ \rho_{21} & 1\end{bmatrix}.$$

So, its inverse is equal to

$$\boldsymbol{R}^{-1}=\frac{1}{|\boldsymbol{R}|}\begin{bmatrix}1-\rho_{12}^2 & \rho_{12}\rho_{2y}-\rho_{1y} & \rho_{1y}\rho_{12}-\rho_{2y}\\ \rho_{2y}\rho_{12}-\rho_{1y} & 1-\rho_{2y}^2 & \rho_{1y}\rho_{2y}-\rho_{12}\\ \rho_{1y}\rho_{12}-\rho_{2y} & \rho_{2y}\rho_{1y}-\rho_{12} & 1-\rho_{1y}^2\end{bmatrix},$$

where $|\boldsymbol{R}|=1-\rho_{1y}^2-\rho_{12}^2-\rho_{2y}^2+2\rho_{1y}\rho_{2y}\rho_{12}$. By derivation with respect to u_1, u_2 from the three-variable Gaussian-copula we obtain

$$D_{12}C^{Ga}(u_1,u_2,v)=\frac{\exp(\frac{-1}{2(1-\rho_{12}^2)}(x_1^2+x_2^2-2\rho_{12}x_2x_1))}{\phi(x_1)\phi(x_2)2\pi(1-\rho_{12}^2)^{\frac{1}{2}}}$$
$$\times\Phi\Bigg(\frac{y-\frac{(\rho_{1y}-\rho_{2y}\rho_{12})x_1+(\rho_{2y}-\rho_{1y}\rho_{12})x_2}{1-\rho_{12}^2}}{\sqrt{\frac{|\boldsymbol{R}|}{1-\rho_{12}^2}}}\Bigg)$$

Also, since X_1, X_2 are related by the Gaussian-copula as

$$C^{Ga}(u_1,u_2)=\int_{-\infty}^{\Phi^{-1}(u_2)}\int_{-\infty}^{\Phi^{-1}(u_1)}\frac{1}{\sqrt{(2\pi)^2|\boldsymbol{R}_{xx}|}}\exp\bigg(\frac{-1}{2}(t_1,t_2)\boldsymbol{R}_{xx}^{-1}(t_1,t_2)^T\bigg)\,dt_1\,dt_2,$$

again its derivation W.R.T u_2 and u_1 yields

$$D_{12}C^{Ga}(u_1,u_2)=\frac{1}{\phi(x_1)\phi(x_2)2\pi(1-\rho_{12}^2)^{\frac{1}{2}}}\exp\bigg(\frac{-1}{2(1-\rho_{12}^2)}(x_1^2+x_2^2-2\rho_{12}x_2x_1)\bigg).$$

Now using the Theorem 2.1, we have

$$
\begin{aligned}
E(Y|X_1=x_1, X_2=x_2) &= \int y \frac{\partial}{\partial y} \frac{D_{12}C(u_1,u_2,v)}{D_{12}(u_1,u_2)}\, dy \\
&= \int y \frac{\partial}{\partial y} \frac{D_{12}C^{Ga}(u_1,u_2,v)}{D_{12}C^{Ga}(u_1,u_2)}\, dy \\
&= \int y \frac{1}{\sqrt{\frac{|\boldsymbol{R}|}{1-\rho_{12}^2}}} \phi\left(\frac{y - \frac{(\rho_{1y}-\rho_{2y}\rho_{12})x_1+(\rho_{2y}-\rho_{1y}\rho_{12})x_2}{1-\rho_{12}^2}}{\sqrt{\frac{|\boldsymbol{R}|}{1-\rho_{12}^2}}} \right) dy \\
&= \frac{\rho_{1y}-\rho_{2y}\rho_{12}}{1-\rho_{12}^2} x_1 + \frac{\rho_{2y}-\rho_{1y}\rho_{12}}{1-\rho_{12}^2} x_2.
\end{aligned}
$$

Detailed computation of Example 2.5. We have the 3-t-copula function as follows:

$$
\begin{aligned}
C^t(u_1,u_2,v) = &\int_{-\infty}^{T_\nu^{-1}(v)} \int_{-\infty}^{T_\nu^{-1}(u_2)} \int_{-\infty}^{T_\nu^{-1}(u_1)} \frac{\Gamma(\frac{\nu+3}{2})|\boldsymbol{R}|^{\frac{-1}{2}}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{-3}{2}}} \\
&\times \left(1 + \frac{(s,t_1,t_2)\boldsymbol{R}^{-1}(s,t_1,t_2)^T}{\nu}\right)^{-\frac{\nu+3}{2}} dt_1\, dt_2\, ds,
\end{aligned}
$$

where $v = F(y)$, $u_1 = G_1(x_1)$ and $u_2 = G_2(x_2)$. Consider the correlation matrix and its partitions are

$$
\boldsymbol{R} = \begin{bmatrix} 1 & \boldsymbol{r}_{yx} \\ \boldsymbol{r}_{yx}^T & \boldsymbol{R}_{xx} \end{bmatrix}, \qquad \mathbf{r}_{yx} = \begin{bmatrix} \rho_{1y} & \rho_{2y} \end{bmatrix}, \qquad \boldsymbol{R}_{xx} = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix}.
$$

So, its inverse is equal to

$$
\boldsymbol{R}^{-1} = \frac{1}{|\boldsymbol{R}|} \begin{bmatrix} 1-\rho_{12}^2 & \rho_{12}\rho_{2y}-\rho_{1y} & \rho_{1y}\rho_{12}-\rho_{2y} \\ \rho_{2y}\rho_{12}-\rho_{1y} & 1-\rho_{2y}^2 & \rho_{1y}\rho_{2y}-\rho_{12} \\ \rho_{1y}\rho_{12}-\rho_{2y} & \rho_{2y}\rho_{1y}-\rho_{12} & 1-\rho_{1y}^2 \end{bmatrix},
$$

where $|\boldsymbol{R}| = 1-\rho_{1y}^2-\rho_{12}^2-\rho_{2y}^2+2\rho_{1y}\rho_{2y}\rho_{12}$. By derivation with respect to u_1, u_2 from the three-variable t-copula we obtain

$$
\begin{aligned}
&D_{12}C^t(u_1,u_2,v) \\
&= \frac{\Gamma(\frac{\nu+3}{2})|R|^{\frac{-1}{2}}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{3}{2}} t_\nu(m(x_1))t_\nu(m(x_2))} \int_{-\infty}^{m(y)} \left(1 + \frac{1-\rho_{12}^2}{\nu|R|}\left(s - \left(\frac{\rho_{1y}-\rho_{2y}\rho_{12}}{1-\rho_{12}^2} m(x_1)\right.\right.\right. \\
&\quad \left.\left.\left. + \frac{\rho_{2y}-\rho_{1y}\rho_{12}}{1-\rho_{12}^2} m(x_2)\right)\right)^2 + \frac{m^2(x_1)+m^2(x_2)-2\rho_{12}m(x_1)m(x_2)}{\nu(1-\rho_{12}^2)}\right)^{\frac{-(\nu+3)}{2}} ds,
\end{aligned}
$$

Now considering that X_1 and X_2 they are connected by t-copula as

$$
C^t(u_1,u_2) = \int_{-\infty}^{T_\nu^{-1}(u_2)} \int_{-\infty}^{T_\nu^{-1}(u_1)} \frac{1}{2\pi\sqrt{1-\rho_{12}^2}} \left(1 + \frac{k_1^2 - 2\rho_{12}k_1k_2 + k_2^2}{\nu(1-\rho_{12}^2)}\right)^{-\frac{\nu+2}{2}} dk_1 dk_2,
$$

where T_ν^{-1} is the inverse of distribution function of t-distribution with ν degrees of freedom. By derivation relative to u_1, u_2 we readily have

$$D_{12}C^t(u_1,u_2) = \frac{1}{t_\nu(m(x_1))t_\nu(m(x_2))2\pi\sqrt{1-\rho_{12}^2}} \times \left(1 + \frac{m^2(x_1) - 2\rho_{12}m(x_1)m(x_2) + m^2(x_2)}{\nu(1-\rho_{12}^2)}\right)^{\frac{-(\nu+2)}{2}}.$$

Again, Theorem 2.1 leads us to

$$\begin{aligned} E(Y|X_1 = x_1, X_2 = x_2) &= \int y \frac{\partial}{\partial y} \frac{D_{12}C(u_1,u_2,v)}{D_{12}C(u_1,u_2)} dy \\ &= \int y \frac{\partial}{\partial y} \frac{D_{12}C^t(u_1,u_2,v)}{D_{12}C^t(u_1,u_2)} dy \\ &= \int y \frac{\partial}{\partial y} \frac{l}{h} dy \end{aligned}$$

which is (11).

Acknowledgments

The authors wish to thank the anonymous reviewers for the comments and suggestions that did lead to a significant improvement of the original manuscript.

Funding

The work of the third author was supported by the project APVV-18-0052 and VEGA 1/0006/19.

References

Acar, E. F., Azimaee, P. and Hoque, Md. E. (2019). Predictive assessment of copula models. *Canadian Journal of Statistics* **47**, 8–26. MR3919892 https://doi.org/10.1002/cjs.11468

Alqawba, M., Diawara, N. and Kim, J. (2019). Copula directional dependence of discrete time series marginals. *Communications in Statistics-Simulation and Computation*, 1–18. MR4336363 https://doi.org/10.1080/03610918.2019.1630434

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning and Inference* **138**, 3616–3633. MR2450101 https://doi.org/10.1016/j.jspi.2005.07.014

Awad, M. and Khanna, R. (2015). Support vector regression. *Efficient learning machines*, 67–80.

Bates, D. M. and Watts, D. G. (2007). *Nonlinear Regression Analysis and Its Applications, Vol. 2*. New York: Wiley. MR1060528 https://doi.org/10.1002/9780470316757

Bennafla, D., Bouchentouf, A., Rabhi, A. and Sabri, Kh. (2016). On the recursive estimation using copula function in the regression model. *New Trends in Mathematical Sciences* **4**, 25. MR3455640 https://doi.org/10.20852/ntmsci.2016115601

Bentler, P. (1985). A new look at the statistical identification model. *IEEE Transactions on Automatic Control* **19**, 716–723.

Chang, B. and Joe, H. (2019). Prediction based on conditional distributions of vine copulas symmetrical linear models. *Computational Statistics & Data Analysis* **139**, 45–63. MR3952615 https://doi.org/10.1016/j.csda.2019.04.015

Crane, G. J. and van der Hoek, J. (2008). Conditional expectation formulae for copulas. *Australian & New Zealand Journal of Statistics* **50**, 53–67. MR2414655 https://doi.org/10.1111/j.1467-842X.2007.00499.x

Cysneiros, F., Cordeiro, G. and Cysneiros, A. (2010). Corrected maximum likelihood estimators in heteroscedastic symmetric nonlinear models. *Journal of Statistical Computation and Simulation* **80**, 451–461. MR2604169 https://doi.org/10.1080/00949650802706420

Cysneiros, F. J. A., Paula, G. A. and Galea, M. (2007). Heteroscedastic symmetrical linear models. *Statistics & Probability Letters* **70**, 1084–1090. MR2395064 https://doi.org/10.1016/j.spl.2007.01.012

Frahm, G., Junker, M. and Szimayer, A. (2003). Elliptical copulas: Applicability and limitations. *Statistics & Probability Letters* **63**, 275–286. MR1986327 https://doi.org/10.1016/S0167-7152(03)00092-0

Hoang, Q., Khandelwal, P. and Ghosh, S. (2019). Robust predictive model using copulas. *Data-Enabled Discovery and Applications* **3**, 8.

Kersting, K., Plagemann, C., Pfaff, P., Burgard, W. L., Krzyżak, A. and Yuille, A. (2007) *Proceedings of the 24th International Conference on Machine Learning*, 393–400.

Kim, Li, S. Y. and Spiegelman, D. (2016). A semiparametric copula method for Cox models with covariate measurement error. *Lifetime Data Analysis* **22**, 16. MR3447180 https://doi.org/10.1007/s10985-014-9315-7

Kole, E., Koedijk, K. and Verbeek, M. (2007). Selecting copulas for risk management. *Journal of Banking & Finance* **31**, 2405–2423.

Konishi, S. (2014). *Introduction to Multivariate Analysis: Linear and Nonlinear Modeling*. Boca Raton: CRC Press. MR3222571

Kumar, P. and Shoukri, M. M. (2007). Copula based prediction models: An application to an aortic regurgitation study. *BMC Medical Research Methodology* **7**, 21.

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. E. and Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical Engineering Online* **6**, 23.

Marsh, L. C. and Cormier, D. R. (2001). Spline regression models. *Sage.*

Masarotto, G., Varin, C., et al (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6**, 1517–1549. MR2988457 https://doi.org/10.1214/12-EJS721

Mesiar, R., Sheikhi, A. and Komorníková, M. (2019). Random noise and perturbation of copulas. *Kybernetika* **55**, 422–434. MR4014595 https://doi.org/10.14736/kyb-2019-2-0422

Noh, H., Ghouch, El, A. and Bouezmarni, T. P. (2013). Copula-based regression estimation and inference. *IEEE Transactions on Automatic Control* **108**, 676–688. MR3174651 https://doi.org/10.1080/01621459.2013.783842

Pitt, M., Chan, D. and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537–554. MR2261441 https://doi.org/10.1093/biomet/93.3.537

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, Th., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., et al. (2015). Package ‘VineCopula’. package version 2.

Seber, G. A. F. and Wild, Ch. J. (2003). *Nonlinear Regression, Vol. 62*, 63. Hoboken. New Jersey: John Wiley & Sons. MR0986070 https://doi.org/10.1002/0471725315

Sheikhi, A. and Mesiar, R. (2020). Copula-based measurement error models. *Iranian Journal of Fuzzy Systems* **17**, 29–38. MR4155893

Wang, C. and Neal, R. M. (2012). Gaussian process regression with heteroscedastic or non-Gaussian residuals. arXiv preprint. Available at 1212.6246. MR3295214

Xu, L., Krzyżak, A. and Yuille, A. (1994). On radial basis function nets and kernel regression: Statistical consistency, convergence rates, and receptive field size. *Neural Networks* **7**, 609–628.