İSTANBUL UNIVERSITY PRESS

EKOIST Journal of Econometrics and Statistics
EKOIST 2023, 39 : 161–171
DOI: 10.26650/ekoist.2023.39.1333281

RESEARCH ARTICLE

A Comparative Perspective on Multivariate Modeling of Insurance Compensation Payments with Regression-Based and Copula-Based Models

Övgücan Karadağ Erdemir[1]

[1](Res. Asst.), Hacettepe University, Faculty of Science, Department of Actuarial Science, Ankara, Türkiye

ABSTRACT

In this study, compensation payments for Turkish motor vehicles' compulsory third-party liability insurance between 2018 and 2022 are modeled from a comparative perspective using regression-based and copula-based multivariate statistical methods. The assumption of gamma distribution for logarithmic compensation payment variables is carried out in both approaches. Bivariate gamma regression is established using the bivariate gamma distribution, and the mixture of experts, one of the machine learning techniques, is employed to form the mixture of bivariate gamma regressions. The bivariate copula regression and finite mixture of copula regression models are designed using the Gumbel and Frank copula functions. The computational analyses were conducted using the mvClaim package in R. Based on the comparison of model results, a mixture of copula-based models is found to be more suitable for the multivariate modeling of insurance compensation payments.

Keywords: Bivariate Gamma Distribution, Copula, Generalized Linear Model, Copula Regression, Insurance Compensation Payments, Machine Learning Techniques, Mixture of Experts Model

1. Introduction

In actuarial science, statistics, econometrics, and financial studies, a multivariate structure is quite common. There may be correlations or dependencies among variables due to multivariability, making identifying, modeling, and incorporating the dependency structure into calculations important to obtain more accurate estimates. The dependence between variables can be determined simply through covariance and correlation analysis. Standard regression models are commonly used to depict the relationship between the response variable and explanatory variables. As the marginal generalized linear model (GLM) represents a generalized form of a linear model, it can be employed with a more diverse range of data as an alternative to linear models. However, apart from the linear regression model and GLM, for modeling correlated multivariate data, multivariate distributions or copulas, which are mathematically based functions, are required.

Random vector variables are utilized in place of a random variable in multivariate distributions. Essential descriptive statistics are summarized using joint probability density and joint cumulative distribution functions. In many multivariate statistical analysis techniques, such as canonical correlation, discriminant analysis, and multivariate analysis of variance, the assumption of a multivariate Gaussian distribution is used (Tatlıdil, 1996). Besides the Gaussian distribution, various other continuous distributions can be used for multivariate modeling. For instance, the bivariate gamma distribution is applied in actuarial science to model joint claim severities (Hu et al., 2019; Hu et al., 2021). A multivariate Pareto distribution is proposed for financial risk measurement (Su and Furman, 2017). Additionally, aside from continuous distributions, discrete distributions can also be adapted into a multivariate form. Vernic (2000) introduced a generalization of the multivariate generalized Poisson distribution. Moreover, in recent times, the phase-type distribution (Zadeh and Bilodeau, 2013; Eryılmaz, 2017) and Sarmanov distribution (Vernic et al., 2022) have frequently appeared in multivariate analysis.

Multivariate modeling has found widespread use in statistics, econometrics, and finance. In actuarial science, there also exist correlated or dependent random variables that necessitate multivariate modeling. In non-life actuarial calculations, model-based approaches are frequently employed. Models are constructed using claim, loss, or risk variables such as claim severity, claim frequency, probability of claim, individual or aggregate losses, deductibles, limits, loss elimination or inflation ratios, value at risk,

Corresponding Author: Övgücan Karadağ Erdemir **E-mail:** ovgucan@hacettepe.edu.tr
Submitted: 26.07.2023 • **Revision Requested:** 10.08.2023 • **Last Revision Received:** 25.08.2023 • **Accepted:** 10.10.2023

and risk exposure, (Klugman et al., 2012). In studies involving claim variables, the assumption of independence among claims has been commonly utilized until recently. However, nowadays, studies that embrace the notion of dependency between claim variables have superseded those relying on the assumption of independence.

Dependency modeling studies in non-life insurance mathematics vary based on the types of claim variables and the structure of the dependency. Generally, the emphasis has been on modeling the dependency between claim severity and frequency; however, different dependency structures have also been under consideration. To model the dependency between claim severity and frequency, various techniques can be employed, such as copulas (Boateng et al., 2017), GLMs (Garrido et al., 2016), GLMMs (Jeong et al., 2017), and copula regression models (Song et al., 2009; Parsa and Klugman, 2011; Czado et al., 2012; Krämer et al., 2013; Maarotto and Varin, 2017; Erdemir and Sucu, 2022). Furthermore, the dependency among claim occurrences (Arvidsson and Francke, 2007), the dependence between types of risks or types of claims (Frees et al., 2010; Ren, 2012), and the dependence between claim severities (Hu et al., 2019; Hu and O'Hagan, 2021) are modeled using various methods.

Machine learning constitutes a subset of artificial intelligence that empowers computer systems to learn, particularly from extensive datasets (Hastie et al., 2009; Alpaydın, 2010; Murphy, 2012). This field finds diverse applications within statistical and financial domains. In statistical studies, machine learning serves such purposes as estimation, regression, classification, and clustering. In financial studies, it proves instrumental for tasks such as risk assessment, portfolio management, stock price estimation, option pricing, and credit evaluation. In actuarial sciences, machine learning techniques are used for a wide range of applications, such as premium estimation, loss estimation, risk management, and reinsurance optimization. In recent years, machine learning techniques have been frequently utilized in actuarial studies that involve claim data (Weerasinghe and Wijegunasekara, 2016; Dewi et al., 2019; Singh et al., 2019; Abdelhadi et al., 2020; Hanafy and Ming, 2021). Moreover, machine learning techniques have been integrated into copula and GLM methods to enhance predictions. For the dependency modeling of multivariate claim severities, a novel approach incorporating a bivariate gamma distribution and a mixture of experts (MoE) has been introduced by Hu et al. (2019). The bivariate gamma MoE model family for joint claim severity is comprised of stages, such as bivariate gamma distribution estimation, mixture of bivariate gamma clustering, bivariate gamma regression, and mixture of bivariate gamma regressions (Hu et al., 2021). MoE is a machine learning technique designed to improve predictive performance through ensemble techniques. The MoE family employs a clustering framework by dividing the problem space into homogeneous regions. While ensemble techniques use results from all models, the MoE family employs results from a few, or only one, expert network(s) (Baldacchino et al., 2016).

Hu and O'Hagan (2021) proposed a new approach named finite mixture of copula regression by integrating copula functions. The copula regression model can be constructed using Gaussian copula function for gamma and Poisson margins under the mixed copula approach (Song, 2007; Czado et al., 2012). Additionally, Archimedean copula functions, such as Gumbel and Frank copulas, can be utilized for copula regression with gamma and zero-truncated Poisson margins (Krämer et al., 2013), and also for the mixture of copula regressions with gamma margins (Hu et al., 2021). Hu et al. (2021) have also introduced a new R package named mvClaim for the multivariate modeling of general insurance claim severities. The mvClaim R package is a recent and valuable resource that offers flexible multivariate modeling of dependency for joint claim severity. Moreover, since it can be adapted to any continuous insurance data, in this study, it is employed in the multivariate modeling of insurance compensation payments.

Compensation payments represent a significant expense for insurance companies. Their accurate modeling and forecasting are vital for determining reserve calculations, estimating future expenses, and establishing budgets for companies. Surprisingly, there is a dearth of work on statistical modeling of compensation payments. When reviewing the literature, it becomes evident that computational calculations are primarily based on fundamental mathematical calculations and legal adjustments, which fall short of true statistical modeling. Notably, the calculations for traffic insurance compensation payments tend to be primarily focused on legal adjustments from the perspective of the insured (Emekliler, 2017; Yolal, 2019).

In this study, compensation payments are regarded as a pivotal expense for insurance companies, with the emphasis placed on modeling using multivariate statistical methods. The proposed multivariate methods and the R package introduced by Hu et al. (2021) for actuarial claim severities have been applied to continuous multivariate traffic insurance compensation payments. Traffic insurance stands as one of the most fundamental legal obligations for all vehicle owners. Insurance companies are obligated to provide compensation payments, such as material, death, invalidity, and medical reimbursements under the umbrella of traffic insurance following the occurrence of a claim.

Compensation payments for motor vehicles' compulsory third-party liability insurance in Turkey are determined according to the General Conditions of Highways Motor Vehicles Compulsory Financial Liability Insurance (https://www.tsb.org.tr/). Material compensation has been computed based on such factors as type of vehicle, market value of the vehicle, and the usage status of the vehicle. This calculation has been facilitated by specific coefficients and a formula outlined in these conditions. Regarding invalidity compensation, both temporary incapacity and permanent disability compensations were calculated in accordance with

the conditions specified in these regulations. These calculations consider the active and passive periods of the individual using life annuities under some actuarial assumptions (Şahin et al., 2021). The death benefit, also recognized as compensation for loss of support, was determined utilizing life tables, actuarial assumptions, and life annuities (Şahin et al., 2020).

In this study, a comparative analysis is conducted regarding the multivariate modeling of the dependency among claim compensation payments between 2018 and 2022. This is achieved through the application of the mixture of bivariate gamma regression using the MoE approach and the finite mixture of copula regression model.

The remainder of the article is organized as follows. Section 2 provides a brief description of the methods employed. This section covers regression-based multivariate models, such as bivariate gamma regression and the mixture bivariate gamma regression models, as well as copula-based multivariate models, like copula regression and the finite mixture of copula regression. In Section 3, an application is performed for the multivariate modeling of Turkish compulsory traffic insurance compensation payments using real data which was obtained from reports about motor vehicle insurance statistics published by the Insurance Association of Turkey, and the results are given. A comparative perspective is presented and concluding remarks are given in Section 4.

2. Methodology

2.1. Regression-Based Multivariate Models

2.1.1. Bivariate Gamma Regression

Consider Y_1 and Y_2 as mixtures of independent gamma-distributed random variables, namely $X_1, X_2 and X_3$, where $X_i \sim Gamma(\alpha_i, \beta)$ for i=1,2,3; $\alpha_i > 0$ and $\beta > 0$. The shape parameters vary for each variable, while the rate parameter remains constant. Let the vector $Y=\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ have a bivariate gamma distribution with the parameters $\alpha_1, \alpha_2, \alpha_3$ and β where $Y_1 = X_1 + X_2$ and $Y_2 = X_2 + X_3$. The probability desnsity function of Y_1 and Y_2 is provided in Eq. 1 (Hu et al., 2021).

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\beta\alpha_1 + \alpha_2 + \alpha_3 \bar{e}\beta(y_1 + y_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \int_{x_3=0}^{min(y_1,y_2)} e^{\beta x_3} x_3^{\alpha_3-1}(y_1 - x_3)^{\alpha_1-1}(y_2 - x_3)^{\alpha_2-1} dx_3 \quad (1)$$

A bivariate gamma regression model which models the relationship between gamma-distributed response variables X_1 and X_2 with a covariate vector z_i^T, can be expressed using logarithmic link function, as shown in Eq. 2. Here, γ_1 and γ_2 represent the coefficients of regression model, and i=1, 2, . . . , n (Purhadi et al., 2018). T represents the transpose of the matrix.

$$\mu_{i1} = E(X_1) = exp(\gamma_1 z_i^T), \mu_{i2} = E(X_2) = exp(\gamma_2 z_i^T) \quad (2)$$

2.1.2. Mixture of Bivariate Gamma Regression

The MoE model is a machine learning method that encompasses model-based clustering with concomitant covariates w_i and provides a more convenient approach for regression modeling. The covariates w_i's are used in the estimation of future outcome variables. Consider a population comprising G components, each characterized by a bivariate gamma distribution with component-specific parameters $\theta_g = (\alpha_{1g}, \alpha_{2g}, \alpha_{3g}, \beta_g)$, where g=1, 2, . . . , G. The conditional density function based on covariates, using the mixing proportion τ_g where $\sum_{g=1}^{G} \tau_g = 1$ is presented in Eq. 3.

$$p(y_i|w_i) = \Sigma_{g=1}^{G} \tau_g(w_{0i}) p\left(y_{1i}, y_{2i}|\alpha_{1ig}(w_{1i}), \alpha_{2ig}(w_{2i}), \alpha_{3ig}(w_{3i})\right) \quad (3)$$

In Eq. 3, $\tau_g(w_{0i})$ represents the gating network and is modeled by multinomial logistic regression as $\widehat{\tau_g}(w_0 i) = \frac{exp(\gamma_{0g}^T w_{0i})}{(\sum_{g}^{'=1} G exp\left(\hat{\gamma}_{0g'}^T w_{0i}\right)}$.

The expert network is denoted by $p\left(y_{1i}, y_{2i}|\alpha_{1ig}(w_{1i}), \alpha_{2ig}(w_{2i}), \alpha_{3ig}(w_{3i})\right)$ and it is modelled using GLM with a logarithmic link function. Specifically, $log(\alpha_{1ig}) = \gamma_{1g}^T w_{1i}, log(\alpha_{2ig}) = \gamma_{2g}^T w_{2i}, log(\alpha_{3ig}) = \gamma_{3g}^T w_{1i}$ and $log(\beta_{ig}) = gamma_{4g}^T w_{4i}$. Here, γ_{0g}, γ_{1g}, γ_{2g}, γ_{3g}, and γ_{4g} are the coefficients of regression models for each component.

Bivariate gamma regression models without mixtures are designed as model types EI and IE, where EI and IE correspond to the bivariate gamma regression over α_{ki} and bivariate gamma regression over β_{ki}, respectively. "E" signifies equal (with a predefined

mixing proportion of 1/G, where G is a parameter associated with gating), while "I" signifies identical density parameters without covariates. In the bivariate gamma regression model (EI), the parameters are α_{ki} and β, while in the bivariate gamma regression model (IE), the parameters are α_k and β_i. Bivariate gamma regression models have no gating network, with G=1 for both models. However, both models include covariates in the expert networks.

A mixture of bivariate gamma regression models has been established based on model types VC, VI, VV, VE, CV, IV, EV, EC, and CE, where "E," "C," and "V" signify equal, constant, and variable, respectively. The bivariate gamma MoE model family employs various parameterizations using "C," "V," and "E." C, V, E, and I are notations established for modeling gating τ tau, expert α_k, k=1,2,3 and expert β parameters with different choices under the MoE approach and are used to facilitate parameterization. $C(\tau_g; \nexists\gamma_{0g})$, $V(\tau_{ig}; \exists\gamma_{0g})$ and $E(\tau_g=1/G)$ are representations of gating τ. $C(\alpha_k g; \nexists\gamma_{kg})$, $V(\alpha_{kig}; \exists\gamma_{kg})$, $E(\alpha_k i; \exists\gamma_k)$ and $I(\alpha_k; \nexists\gamma_k)$ are for expert α_k, while $C(\beta_g; \nexists\gamma_{4g})$, $V(\beta_i g; \exists\gamma_{4g})$, $E(; \exists\gamma_4)$ and $I(\beta; \nexists\gamma_4)$ are for expert β. All mixture bivariate gamma regression models have gating networks and include covariates in the expert networks. In mixture models, data is segmented into specific clusters using machine learning techniques, such as MoE. For more comprehensive information, refer to Hu et al. (2019) and Hu et al. (2021).

2.2. Copula-Based Multivariate Models

2.2.1. Copula Regression

A parametric copula function is a useful multivariate distribution function, as represented by Eq. 4. The marginal distribution functions $F_{X_1}(x_1), F_{X_2}(x_2), \ldots., F_{X_n}(x_n)$, are uniformly distributed within the interval [0,1], with θ representing the copula parameter, as per Sklar's Theorem (Sklar, 1959; Nelsen, 2007).

$$F_{X_{(1,\ldots X_n)}}(x_1, \ldots, x_n) = C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)|\theta) \tag{4}$$

Copulas, which are frequently preferred in multivariate and dependency modeling, can either be used independently or be incorporated into copula regression models through the inclusion of GLM techniques (Czado et al., 2012; Kramer et al., 2013; Masarotto and Varin, 2017; Erdemir and Sucu, 2022). In addition to understanding the distribution that variables conform to, the utilization of copulas in copula regression models requires the incorporation of certain covariates. Copula regression models can be defined using the Gaussian copula function within the mixed copula approach (Song, 2007; Song et al., 2009; Czado et al., 2012). Kramer et al. (2013) also employed Archimedean copulas in their copula regression models. A bivariate copula regression model can be constructed as C(Gamma GLM,Gamma GLM|θ), combining marginal Gamma GLMs through a C(.,.|θ) Archimedan copula function with θ as the copula parameter.

2.2.2. Finite Mixture of Copula Regressions

Hu and O'Hagan (2021) defined a finite mixture of copula regression models that encompass Joe, Gumbel, Clayton, Frank, and survival Clayton copulas, in addition to the Gaussian copula. Similar to the mixture of bivariate gamma regression, the MoE-based modeling includes a combination of two copula regression models. A GLM is expressed using δ as the link function and β_{jg} as the regression coefficient for the j^{th} margin and g^{th} component, as shown in Eq. 5. Additionally, x_{ijg}^T represents the covariate of j^{th} margin and g^{th} component, where j=1,2; g=1, 2, . . . , G.

$$\mu_{ji} = \delta^{-1}\left(x_{ijg}^T \beta_{jg}\right) \tag{5}$$

The likelihood function $L(\theta) = \prod_{i=1})^N \sum_{g=1}^G \tau_g h_g(y_i; \theta_g)$ for the finite mixture copula regression model is provided in Eq. 6, where c_g is the density of copula function. The term $h_g(y_i; \theta_g)$ represents the component-specific copula regression for $\theta_g = \{\alpha_g, \beta_{jg}, \gamma_{jg}\}$, with α_g as the copula parameter. An expectation-maximization (EM) algorithm is employed for the parameter estimation (Dempster et al., 1977; Hu et al., 2021). The parameter τ_g signifies the mixing proportion, and it holds that $\sum_{g=1}^G \tau_g = 1$ under the MoE clustering approach.

$$L(\theta) = \prod_{i=1}^{N} \Sigma_{g=1}^G \tau_g c_g(F_1(y_{1i}; \beta_{1g}, \gamma_{1g}), F_2(y_{2i}; \beta_{2g}, \gamma_{2g}); \alpha_g) f_1(y_{1i}; \beta_{1g}, \gamma_{1g}), f_2(y_{2i}; \beta_{2g}, \gamma_{2g}) \tag{6}$$

3. Application

An application of these models have been conducted for the multivariate modeling of Turkish compulsory traffic insurance compensation payments using data extracted from reports that include motor vehicle insurance statistics published on the website of the Insurance Association of Turkey (https://www.tsb.org.tr/tr/istatistikler). The study discusses material, death, and invalidity compensation amounts paid within the scope of compulsory traffic insurance for different vehicle types on a quarterly basis from 2018 to 2022. Medical, outstanding, and transferring outstanding compensation payments have not been included in the study. The data in the reports is not used directly. For all three types of compensation, the final compensation amounts are calculated based on the reports as (Compensation Payments to the Insured) + (Expert Payments) + (Other Cost Payments) = Total Compensation. Material, death, and invalidity compensation payments are determined as response variables, while the premium, number of benefits, term, and type of vehicle are considered covariates. To prepare the data for modeling, it was categorized according to the type of vehicle and the term. The term was categorized into four quarters (1st, 2nd, 3rd, and 4th), and the type of vehicle was categorized into two groups: automobiles and non-automobiles. The type of vehicle variable is represented as a dummy variable (1, 0). The other covariates are treated as continuous variables. Ultimately, a dataset with 300 observations, four covariates, and three response variables has been compiled. One of the reasons for transforming the data into categorical values is the utilization of GLM in multivariate models. Furthermore, due to the nature of the insurance system, it employs more categorical data based on the characteristics of the policyholder or the vehicle, rather than individual data. The categorical structure facilitates interpretation and estimation.

The aim of this study is to achieve the multivariate modeling of compensation payments based on vehicle type, premium, number of benefits, and term variables using both regression-based and copula-based models. The regression-based models encompass bivariate gamma regression models, as well as mixtures of these regression models created through machine learning techniques. The copula-based models include copula regression models and mixtures of copula regression models utilizing Gumbel and Frank copula functions. Statistical analyses and actuarial calculations were primarily conducted using the mvClaim R package (Hu et al., 2021). The methods were applied using BGR ( ), MBGR ( ), MCGR ( ), and copreg.gamma ( ) functions in the mvClaim package. The processing steps in functions are given in detail in the research of Hu et al. (2021). Additionally, auxiliary packages including copula, stats, lme4, ggplot2, PerformanceAnalytics, and fitdistrplus have also been utilized.

Before conducting multivariate modeling, the correlation between Turkish compulsory traffic insurance compensation payments was assessed through the correlation matrix. The correlations between material-death, material-invalidity, and death-invalidity was found to be 0.940109, 0.957997, and 0.962905, respectively. These correlation coefficients are visualized in Figure 1. The study demonstrates a significant and high dependency between the selected compensation payments, indicating the potential for multivariate modeling. The compulsory traffic insurance covers the compensations to be paid for the reparation of the damage that arises as a result of an event such as a traffic accident. The compensations may depend on certain factors related to the general structure of traffic accidents and the functioning of the insurance system. In traffic accidents that result in death or injury, the occurrence of material damage is inevitable. Therefore, a high correlation between these compensation payments is expected among these payment types.

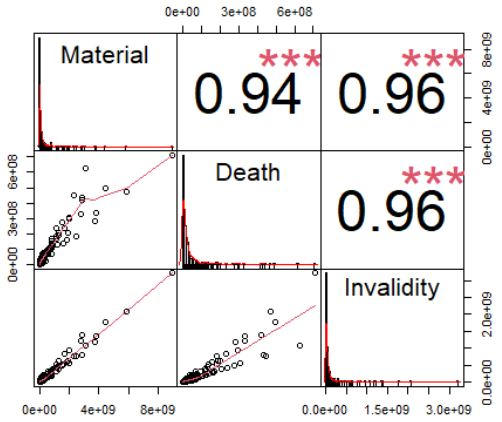

Figure 1. Correlation between death, material and invalidity compensation payments

In this study, Turkish compulsory traffic insurance compensations have been modeled using regression-based and copula-based multivariate methods, with the assumption that compensation variables follow a Gamma distribution. The conditions for the

application of the functions require the data to follow a gamma distribution and possess a structure suitable for multivariate modeling. Thus, the suitability of compensation payment variables to the Gamma distribution has been tested. To fulfill the assumption of a Gamma distribution, a logarithmic transformation was applied to all compensation variables. The shape and rate parameters were determined using the fitdistrplus R package for the logarithmic compensations. According to the Kolmogorov-Smirnov test, the logarithmic death, logarithmic material, and logarithmic invalidity compensations exhibit a good fit to the Gamma distribution ($p>0.05$). Descriptive statistics of logarithmic material, death, and invalidity compensation payments are provided in Table 1 below.

Table 1. The descriptive statistics of log- material, death and invalidity compensation payments

	Material	Death	Invalidity
Minimum	6.7690	5.4640	4.1730
1st Quartile	15.992	14.640	15.096
Median	17.650	16.291	17.029
Mean	17.114	15.589	16.240
3rd Quartile	18.920	17.354	18.123
Maximum	22.913	20.381	21.892
Standard Deviation	2.9924	2.9195	3.1863
Variance	8.9543	8.5233	10.152

Parametric bootstrap-based goodness-of-fit tests were conducted for the elliptical copulas (Gaussian and t) and several Archimedean copulas (Gumbel, Frank, Clayton, and Joe) to model the dependence between material and death compensations, the dependence between material and invalidity compensations, as well as the dependence between death and invalidity compensations separately. The results of these tests are presented in Table 2.

Table 2. The results of goodness-of-fit test of parametric copulas

Copula for the Dependence Between Material-Death Compensations			
Parametric copulas	**Parameter estimation**	**Statistic**	**p-value**
Gaussian	0.94631	0.04557	0.009804*
t	0.94753, 6.07160 #	0.08757	0.049020*
Gumbel	5.0114	0.03677	0.009804*
Frank	16.857	0.10058	0.009804*
Clayton	3.8013	0.39174	0.009804*
Joe	7.0723	0.20793	0.009804*
Copula for the Dependence Between Material-Invalidity Compensations			
Parametric copulas	**Parameter estimation**	**Statistic**	**p-value**
Gaussian	0.96002	0.02583	0.009804*
t	0.96079, 4.75585 #	0.07648	0.068630*
Gumbel	5.6155	0.03309	0.009804*
Frank	19.981	0.10368	0.009804*
Clayton	5.0228	0.21816	0.009804*
Joe	7.5368	0.41273	0.009804*
Copula for the Dependence Between Death-Invalidity Compensations			
Parametric copulas	**Parameter estimation**	**Statistic**	**p-value**
Gaussian	0.95946	0.023055	0.009804*
t	0.96077, 14.72332 #	0.098977	0.009804*
Gumbel	5.2451	0.040724	0.009804*
Frank	20.889	0.145360	0.009804*
Clayton	4.9203	0.214590	0.009804*
Joe	6.6707	0.390200	0.009804*

*P value significant at the 0.05 level, #t copula has two parameters, while the other copulas have only one parameter

According to Table 1, all chosen parametric copulas are suitable for jointly modeling material, death, and invalidity compensations, considering the dependence between compensations ($p<0.05$). Given the strong correlation between the compensations, it is an anticipated outcome that all copulas are well-suited. The suitability of both elliptical and Archimedean copulas for the data

could stem from the presence of multiple dependencies in the data or indicate that the dataset simultaneously represents specific dependency structures from different perspectives. Considering that the analyses are conducted using the mvClaim R package, which is specifically designed for Gumbel and Frank copulas, this study focuses on these two copulas for the multivariate modeling of compensation payments.

Regression-based and copula-based multivariate models were explored for modeling Turkish compulsory traffic insurance compensation payments. Initially, comparisons were conducted within the regression-based models and separately within the copula-based models. Subsequently, the chosen regression-based and copula-based models were compared with each other. Notably, while the mvClaim R package has been primarily designed for insurance claim severities modeling, it can be adapted for positively continuous data, as emphasized by Hu et al. (2021). This study employs a multivariate approach to model continuous compensation payments using a combination of regression models, machine learning techniques, and copula functions.

The bivariate gamma regression and mixture of bivariate gamma regression models are fitted using the EM algorithm through the BGR ( ) and MBGR ( ) functions within the package, respectively. The outcomes of the regression models are consolidated in Table 3 below.

Table 3. Comparison of Bivariate Gamma Regression and Mixture of Bivariate Gamma Regression Models

Bivariate Gamma Regression Models for Material-Death Compensations			
Models	**log-likelihood**	**AIC**	**BIC**
BGR[1] (EI)	-928.6031	**1879.206**	**1919.948**
BGR (IE)	-1032.769	2079.537	2105.464
MBGR[2] (VC)	-752.3417	1556.683	1652.982
MBGR (VI)	-757.0442	1564.088	1656.683
MBGR (VV)	-758.8661	1581.732	1700.253
MBGR (VE)	-727.5208	**1511.042**	**1614.747**
MBGR (CV)	-800.6102	1637.220	1703.889
MBGR (IV)	-1010.359	2050.718	2106.275
MBGR (EV)	-808.2121	1660.424	1741.907
MBGR (EC)	-922.3294	1876.659	1935.919
MBGR (CE)	-841.3727	1710.745	1762.598
Bivariate Gamma Regression Models for Material-Invalidity Compensations			
Models	**log-likelihood**	**AIC**	**BIC**
BGR (EI)	-933.2306	**1888.461**	**1929.203**
BGR (IE)	-1071.726	2157.452	2183.379
MBGR (VC)	-727.1016	**1506.203**	**1602.502**
MBGR (VI)	-766.8502	1583.700	1676.295
MBGR (VV)	-739.5443	1543.089	1661.610
MBGR (VE)	-735.2582	1526.516	1630.222
MBGR (CV)	-772.0130	1580.026	1646.694
MBGR (IV)	-1049.915	2129.831	2185.388
MBGR (EV)	-801.7514	1647.503	1728.986
MBGR (EC)	-921.4530	1874.906	1934.167
MBGR (CE)	-834.7204	1697.441	1749.294
Bivariate Gamma Regression Models for Death-Invalidity Compensations			
Models	**log-likelihood**	**AIC**	**BIC**
BGR (EI)	-955.6750	**1933.35**	**1974.092**
BGR (IE)	-1097.225	2208.45	2234.377
MBGR (VC)	-729.8655	**1511.731**	**1608.029**
MBGR (VI)	-772.4224	1594.845	1687.439
MBGR (VV)	-727.0960	1518.192	1636.713
MBGR (VE)	-735.8603	1527.721	1631.427
MBGR (CV)	-745.2686	1526.537	1593.205
MBGR (IV)	-1087.455	2204.910	2260.467
MBGR (EV)	-801.0448	1646.090	1727.573
MBGR (EC)	-944.0732	1920.146	1979.407
MBGR (CE)	-828.6652	1685.330	1737.183

[1]BGR: Bivariate Gamma Regression, [2]MBGR: Mixture Bivariate Gamma Regression

According to Table 3, among the bivariate regression models, the bivariate gamma regression model (EI) exhibits the lowest values of information criteria (AIC, BIC) for all three pairs of compensation payments. Among the mixture regression models, those with the lowest AIC and BIC values are the mixture of bivariate gamma regression models (VE) (AIC=1511.042, BIC=1614.747), (VC) (AIC=1506.203, BIC=1602.502), and (VC) (AIC=1511.731, BIC=1608.029) for the pairs of material-death compensations, material-invalidity compensations, and death-invalidity compensations, respectively. It is noteworthy that the mixture regression models display lower values of information criteria compared to the bivariate gamma regression models for all three pairs of compensations.

Based on the chosen models using information criteria, Figures 2, 3, and 4 depict the graphs illustrating the estimated logarithmic and actual logarithmic compensation payments for all three pairs. In these figures, "BGR" denotes the bivariate gamma regression model, while "MBGR" represents the mixture of bivariate gamma regression model.

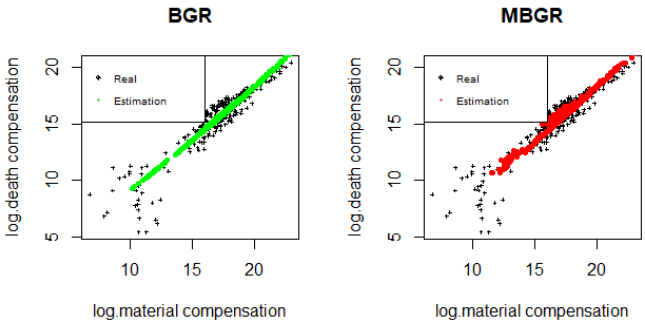

Figure 2. The fitted values of BGR(EI) and MBGR(VE) models for the pair material-death compensations

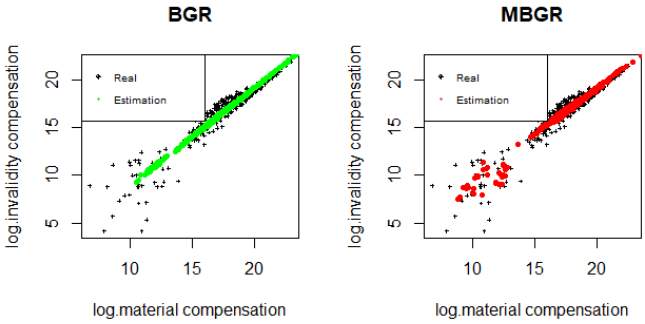

Figure 3. The fitted values of BGR(EI) and MBGR(VC) models for the pair material-invalidity and death-invalidity compensations

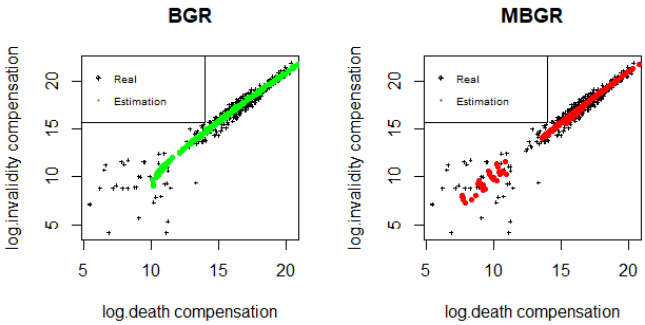

Figure 4. The fitted values of BGR(EI) and MBGR(VC) models for the pair death-invalidity compensations

It has been observed that the predicted values in the mixed bivariate gamma regression models are closer to the actual values when compared to the bivariate gamma regression models for all three pairs of compensations.

The outcomes of bivariate copula regressions with gamma margins, including a single copula (Gumbel or Frank), and the mixture of copula regressions with gamma margins, involving both copulas (Gumbel and Frank), fitted using the EM algorithm, are presented in Table 4. According to Table 4, the mixture of bivariate copula regression models with Gumbel and Frank copulas exhibit lower information criteria in comparison to the copula regression models for all pairs of compensations.

Table 4. Comparison of Copula Regression and Mixture of Copula Regression Models

Copula-Based Models for Material-Death Compensations			
Models	**log-likelihood**	**AIC**	**BIC**
Bivariate Copula Regression with Gumbel Copula	-956.6117	1931.223	1964.557
Bivariate Copula Regression with Frank Copula	-924.8574	1867.715	1901.049
Mix. of Copula Regression with Gumbel and Frank Copulas	-569.9331	**1169.866**	**1225.423**
Copula-Based Models for Material-Invalidity Compensations			
Models	**log-likelihood**	**AIC**	**BIC**
Bivariate Copula Regression with Gumbel Copula	-965.4639	1948.928	1982.262
Bivariate Copula Regression with Frank Copula	-1075.104	2168.208	2201.542
Mix. of Copula Regression with Gumbel and Frank Copulas	-578.0628	**1186.126**	**1241.682**
Copula-Based Models for Death-Invalidity Compensations			
Models	**log-likelihood**	**AIC**	**BIC**
Bivariate Copula Regression with Gumbel Copula	-1020.234	2058.468	2091.802
Bivariate Copula Regression with Frank Copula	-1005.715	2029.429	2062.763
Mix. of Copula Regression with Gumbel and Frank Copulas	-602.2227	**1234.445**	**1290.002**

Finally, upon comparing the selected mixture multivariate models, it becomes evident that the models with copulas are more suitable for the multivariate modeling of this data. As can be seen in Table 5 below, the AIC and BIC values for material-death compensations are calculated as 1169.866 and 1225.423, respectively. The information criteria (AIC, BIC) values for material-invalidity compensations are (AIC=1186.126, BIC=1241.682), and for death-invalidity compensations, they are (AIC=1234.445, BIC=1290.002). The mixture of copula regression models proves to be more suitable, with the criteria for these models being indicated in bold.

Table 5. Comparison of Mixture Multivariate Models

Multivariate Models for Material-Death Compensations		
	Mixture of Bivariate Gamma Regression	**Mixture of Copula Regression**
AIC	1511.042	**1169.866**
BIC	1614.747	**1225.423**
Multivariate Models for Material-Invalidity Compensations		
	Mixture of Bivariate Gamma Regression	**Mixture of Copula Regression**
AIC	1506.203	**1186.126**
BIC	1602.502	**1241.682**
Multivariate Models for Death-Invalidity Compensations		
	Mixture of Bivariate Gamma Regression	**Mixture of Copula Regression**
AIC	1511.731	**1234.445**
BIC	1608.029	**1290.002**

4. Conclusion

Motor vehicles' compulsory third-party liability insurance compensation payments are essential expense items for insurance companies, as traffic insurance is a mandatory liability insurance frequently chosen by policyholders. Calculations of compensation payments primarily rely on fundamental mathematical methods and legal regulations. Predicting compensation amounts is crucial for determining reserve calculations, estimating future expenses, and establishing company budgets. Additionally, there might exist correlations or dependencies among traffic insurance compensation payment variables, given that insurance companies are obligated to provide compensation for material, death, invalidity, and medical claims. Identifying and modeling these dependencies is essential, making multivariate statistical methods the foundation for such calculations.

In non-life insurance mathematics, various approaches, including copula, GLM, GLMM, copula regression models, and multivariate distributions, are employed for dependency modeling studies. Notably, in recent years, machine learning techniques integrated with copula and GLM have gained traction in actuarial studies, significantly enhancing prediction accuracy. In this study, Turkish motor vehicles' compulsory third-party liability insurance compensation payments spanning the years 2018 and 2022 were modeled using regression-based and copula-based multivariate statistical methods. Both models assume the gamma distribution for logarithmic compensation payment variables. Bivariate gamma regression leverages the bivariate gamma distribution, while the mixture of bivariate gamma regressions was realized through the MoE approach, one of the machine learning techniques. The bivariate copula regression and finite mixture of copula regression models were formulated using Gumbel and Frank copula functions. The computational analysis was facilitated using the R package "mvClaim."

MoE was applied with the help of the MBGR ( ) and MCGR functions in the mvClaim package. The primary purpose of MoE is to provide an approach where different experts solve various sub-problems, with a better outcome being achieved by combining the results of these experts. The advantage of MoE is evident in this study, as it leads to more accurate predictions with the mixture models employing the MoE approach.

A comparative approach is presented through information criteria. The model results indicate that the mixture of models, both in regression-based and copula-based scenarios, yields superior outcomes for the multivariate modeling of insurance compensation payments. The observed high correlation between insurance compensation pairs validates the suitability of copula-based models over regression models. The incorporation of machine learning techniques, like MoE, enhances predictions and results in lower information criteria via mixture copula regression models. The primary focus of this study is estimating compensation payments using multivariate statistical methods. For a more comprehensive study, the results obtained from classical compensation calculation methods can be contrasted with the findings of regression-based and copula-based mixture models proposed in this study.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The author have no conflict of interest to declare.

Grant Support: The author declared that this study has received no financial support.

ORCID:

Övgücan Karadağ Erdemir 0000-0002-4725-3588

REFERENCES

Abdelhadi, S., Elbahnasy, K. & Abdelsalam, M. (2020). A Proposed Model to Predict Auto Insurance Claims Using Machine Learning Techniques. *Journal of Theoretical and Applied Information Technology, 98*(22).

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Arvidsson, H. & Francke, S. (2007). Dependence in Non-Life Insurance. *UUDM Project Report.* 0

Baldacchino, T., Cross, E.J., Worden, K. & Rowson, J. (2016). Variational Bayesian Mixture of Experts Models and Sensitivity Analysis for Nonlinear Dynamical Systems. *Mechanical Systems and Signal Processing, 66*, 178–200.

Boateng, M.A., Omari-Sasu, A.Y., Avuglah, R.K. & Frempong, N.K. (2017). On Two Random Variables and Archimedean Copulas. International Journal of Statistics and Applications, 7(4), 228.

Czado, C., Kastenmeier, R., Brechmann E.C. & Min, A. (2012). A Mixed Copula Model for Insurance Claims and Claim Sizes. *Scandinavian Actuarial Journal, 4*, 278.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological), 39*, 1–38.

Dewi, K.C., Murfi, H. & Abdullah, S. (2019). Analysis Accuracy of Random Forest Model for Big Data–A Case Study of Claim Severity Prediction in Car Insurance. *5th International Conference on Science in Information Technology (ICSITech)*, 60-65.

Emekliler, N.A. (2017). Karayolları Motorlu Araçlar Zorunlu Mali Sorumluluk Sigortasında Hasar Oranlarının Hesaplanması ve Hasar Oranlarının Tahmini Emekliler Sigorta Örneği. *Master Dissertation in Turkish, Başkent Üniversitesi Sosyal Bilimler Enstitüsü,* Turkey.

Erdemir, Ö.K. & Sucu, M. (2022). A Modified Pseudo-Copula Regression Model for Risk Groups with Various Dependency Levels. *Journal of Statistical Computation and Simulation, 92*(5), 1092-1112.

Eryılmaz, S. (2017). On Compound Sums Under Dependency. *Insurance: Mathematics and Economics, 72*, 228.

Frees, E.W., Myers G. & David, C. (2010). Dependent Multi-peril Ratemaking Models. *ASTIN Bulletin*, 40, 699.

Garrido, J., Genest, C. & Schulz, J. (2016). Generalized Linear Models for Dependent Frequency and Severity of Insurance Claims. *Insurance: Mathematics and Economics, 70*, 205.

Hanafy, M. & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. *Risks, 9*(2), 42.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Hu, S., Murphy, T.B. & O'Hagan, A. (2019). Bivariate Gamma Mixture of Experts Models for Joint Insurance Claims Modelling. *Cornell University*, arXiv: arxiv.org/abs/1904.04699.

Hu, S., Murphy, T.B. & O'Hagan, A. (2021). MvClaim: An R Package for Multivariate General Insurance Claims Severity Modelling. *Annals of Actuarial Science, 15*(2), 441-457.

Jeong, H., Valdez, E. A., Ahn, J. Y. & Park, S. (2017). Generalized Linear Mixed Models for Dependent Compound Risk Models. SSRN 3045360.

Klugman, S.A., Panjer, H.H. & Willmot, G.E. (2012). Loss models: from data to decisions. *John Wiley & Sons*, 715.

Krämer, N., Brechmann, E.C., Silvestrini, D. & Czado, C. (2013). Total Loss Estimation Using Copula-Based Regression Models. *Insurance: Mathematics and Economics, 53*, 829.

Masarotto, G. & Varin, C. (2017). Gaussian Copula Regression in R. *Journal of Statistical Software, 77* (8).

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Nelsen, R.B. (2007). An Introduction to Copulas. *Springer science & business media.*

Parsa R.A. & Klugman, S.A. (2011). Copula Regression. *Variance Advancing and Science of Risks*, 5, 45.

Purhadi, B. & Purnami, S. (2018). Parameter Estimation and Statistical Test in Bivariate Gamma Regression Model. *8th Annual Basic Science Onternational Conference*.

Ren, J. (2012). A Multivariate Aggregate Loss Model. *Insurance: Mathematics and Economics*, 51, 402.

Singh, R., Ayyar, M.P., Pavan, T.V.S., Gosain, S. & Shah, R.R. (2019). Automating Car Insurance Claims Using Deep Learning Techniques. *IEEE fifth international conference on multimedia big data (BigMM)*, 199-207.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris,* 8, 229–231.

Song, P.X.K (2007). Correlated Data Analysis: Modeling, Analytics, And Applications. *Springer Science & Business Media*, Ontario, Canada.

Song, P.X.-K., Li, M. & Yuan, Y. (2009). Joint Regression Analysis of Correlated Data Using Gaussian Copulas. *Biometrics, 65*(1), 60-68.

Su, J. & Furman, E. (2017). A Form of Multivariate Pareto Distribution with Applications to Financial Risk Measurement. *ASTIN Bulletin: The Journal of the IAA, 47*(1), 331-357.

Şahin, Ş., Nevruz, E., Karageyik, B. B., & Simsek, G. (2020). Destekten Yoksun Kalma Tazminatı Hesaplama Yöntemleri, Şeçkin Yayıncılık, Türkiye.

Şahin, Ş., Karageyik, B. B., Nevruz, E., & Simsek, G. (2021). Aktüerya Bilirkişiliği-İş Göremezlik Tazminatı Hesaplama Yöntemleri, Şeçkin Yayıncılık, Türkiye.

Tatlidil, H. (1996). Uygulamalı Çok Degiskenli İstatistiksel Analiz. *Cem Web Ofset Ltd. Sti,* Ankara.

Vernic, R. (2000). A Multivariate Generalization of The Generalized Poisson Distribution. *ASTIN Bulletin, 30*(1), 57–67.

Vernic, R., Bolancé, C. & Alemany, R. (2022). Sarmanov Distribution for Modeling Dependence Between the Frequency and the Average Severity of Insurance Claims. *Insurance: Mathematics and Economics*, 102, 111-125.

Weerasinghe, K.P.M.L.P. & Wijegunasekara, M.C. (2016). A Comparative Study of Data Mining Algorithms in The Prediction of Auto Insurance Claims. *European International Journal of Science and Technology, 5*(1), 47-54.

Yolal, H.E. (2019). Karayolları Motorlu Araçlar Zorunlu Mali Sorumluluk (Trafik) Sigortalarında Sigorta Tazminatının Ödenmesinde Kusurun Etkisi. *ProQuest Dissertations & Theses Global*.

Zadeh, A. H. & Bilodeau, M. (2013). Fitting Bivariate Losses with Phase-Type Distributions. *Scandinavian Actuarial Journal*, 4, 241.

https://www.tsb.org.tr/tr/istatistikler Access date: 06.04.2023

https://www.tsb.org.tr/media/attachments/Trafik_Genel_%C5%9Eartlar%C4%B1_06122021__Ekler_Dahil.pdf,Trafik Sigortası Genel Şartları, Access date: 21.08.2023

How cite this article

Karadag Erdemir, O. (2023). A comparative perspective on multivariate modeling of insurance compensation payments with regression-based and copula-based models. *EKOIST Journal of Econometrics and Statistics*, 39, 161-171.
https://doi.org/10.26650/ekoist.2023.39.1333281