doi: 10.21608/esju.2024.311618.1039

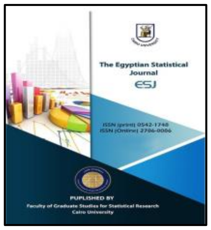
Research Article

The Egyptian Statistical Journal (ESJ),68 (2): 65-85

Homepage: https://esju.journals.ekb.eg/

The Egyptian Statistical Journal

Print ISSN 0542-1748– Online ISSN 2786-0086


The Egyptian Statistical Journal
ESJ
ISSN (print) 0542-1748
ISSN (Online) 2786-0086
PUBLISHED BY
Faculty of Graduate Studies for Statistical Research
Cairo University

Using Tail Dependence on Copula-based Regression Models in Mixed Data

Fatma Y. Alshenawy[1]

Received 11 August 2024; Revise 17 September 2024; accepted 23 September 2024

Keywords

Tail Dependence, Copula-Based Regression Models, Health Insurance Data, Mixed Data, Gaussian Copula, Clayton Copula, Gumbel Copula.

Abstract

This paper explores the efficacy of incorporating tail dependence into copula-based regression models applied to mixed health insurance data. Recognizing the limitations of traditional Generalized Linear Models (GLMs) in capturing the nuanced relationships within mixed data types, we extend the GLM framework to include bivariate and multivariate structures with gamma and negative binomial distributions. We apply a comprehensive suite of copula families—Gaussian, Clayton, Gumbel, Frank, and Student's-t to model the dependencies between variables, focusing on capturing tail dependence, a critical aspect in the context of insurance claim sizes and frequencies. Our methodology involves fitting bivariate GLMs for each pair of variables to understand pairwise dependencies and then extending the analysis to multivariate GLMs to capture the complex interplay between multiple predictors and the response variable. The analysis is performed on a rich dataset of health insurance claims, to identify the copula family that best represents the dependence structure. The results demonstrate that copulas with heavier tails, such as the Gumbel and Student's t copulas, provide superior fit and predictive performance for extreme claim amounts, outperforming those with lighter tails, such as the Gaussian and Frank copulas. The Clayton copula also shows promise in modeling lower tail dependence. Our findings suggest that tail dependence is a significant factor in accurately modeling health insurance claims data, and that the choice of copula family has a profound impact on the model's effectiveness. We conclude that copula-based regression models, with a focus on tail dependence, offer a robust alternative to conventional regression techniques, enabling actuaries and data analysts in health insurance to better understand risk and price policies more accurately. Our research contributes to the actuarial field by providing a systematic comparison of copula families in the context of health insurance data and by underscoring the importance of tail dependence in actuarial modeling

1. Introduction

Health insurance companies are increasingly confronted with heterogeneous data sets, comprising various risk factors that include both typical and extreme values. Traditional regression models, while useful, often fall short in accurately capturing the complex dependencies between these mixed types of data-especially in the tails of the distribution where the most significant insurance claims reside. This inadequacy has driven the development and application of more sophisticated models, such as tail-dependence copula-based regression models, to accurately assess and predict insurance risk.

✉ Corresponding author*: felshinawy@gmail.com

[1]Faculty of Commerce, Mansoura University, Egypt.

The present paper seeks to contribute to the existing body of knowledge by applying a tail-dependence copula-based regression model to mixed health insurance data. The objective is to enhance the understanding of the dependency structure between different types of variables—particularly focusing on the tail dependence that can lead to significant financial impacts due to the occurrence of high-cost claims.

This paper will provide a comprehensive review of the theoretical underpinnings of copula-based regression models, detail the methodology of applying these models to health insurance data, and present a novel empirical analysis. The study aims to offer insights that can aid insurers in risk assessment and policy design, ultimately leading to more robust and financially sustainable health insurance products.

The analysis of mixed data, comprising continuous and discrete variables, is a ubiquitous challenge in statistical modeling. Traditional regression models often prove inadequate in capturing the intricate dependence structures within such datasets, particularly in the tails of the distribution. This limitation is particularly acute in sectors like health insurance, where the occurrence of extreme events—though rare—can significantly impact the system.

The present paper aims to delve into the use of tail-dependence copula-based regression models to better understand and model the dependencies in mixed data, with a particular focus on applications within health insurance data. This approach is poised to offer a more nuanced understanding of the interactions between different types of variables, especially in the context of extreme values.

2. Literature Review

In the realm of statistical analysis, mixed data types pose a substantial challenge in modeling the relationships between variables (Kolev and Paiva, 2009), particularly when considering the joint distribution of such variables. The application of copula-based models has gained traction as a robust means to address this challenge, especially in fields where tail dependence is a critical concern, such as finance and insurance.

A copula is a mathematical function that allows for the modeling of complex dependencies between random variables with different marginal distributions (Nelsen, 2006). The seminal work by Sklar, 1959 introduced the copula function, fundamentally changing the approach to multivariate distribution modeling by allowing for the separation of marginal distributions from their dependence structure.

The importance of capturing tail dependence through copulas in mixed data scenarios is well-documented in the literature. Tail dependence refers to the probability of extreme values in one variable occurring simultaneously with extreme values in another variable. This concept is particularly relevant in risk management (McNeil et al.,2015) and (Anderson, 2012) where the underestimation of joint extreme events can lead to significant financial implications (Embrechts, McNeil, and Straumann, 2002).

In the context of mixed data, copulas have been utilized to model the dependence structure between continuous and discrete variables. Genest and Favre, 2007 explored the use of copulas to understand the relationships between mixed variables, highlighting the flexibility of copulas in accommodating different types of data. The work by Patton, 2012 further advanced the application of copulas in econometrics, providing robust methods for modeling and inference for copula-based models in mixed-data environments.

Recent advancements in copula research have focused on the use of vine copulas, which provide a method for constructing high-dimensional copulas from bivariate copulas, allowing for greater flexibility in modeling complex dependencies (Aas et al., 2009). Joe (2014) provided comprehensive insights into the theory and applications of copulas in multivariate problems, including those involving mixed data.

The literature indicates a growing consensus on the efficacy of tail-dependence copula-based models in capturing the complex interactions within mixed data sets and providing a more accurate understanding of the underlying risk structure.

The complexity of health insurance data, with its mix of continuous, discrete, and categorical variables, has necessitated the development of sophisticated statistical methods to understand and predict the relationships within the data. A growing body of literature has focused on the integration of copula-based models to tackle these challenges, particularly emphasizing the importance of capturing tail dependencies.

Initially, copula models were employed in finance to model the dependency structure between assets, but they have since been adapted for use in insurance data (Embrechts, McNeil, and Straumann, 2002). Sklar,1959 laid the theoretical groundwork for using copulas to model multivariate distributions, allowing for the separation of marginal distributions from their dependency structure (Genest et al., 2007). This separation is particularly useful in mixed data types found in health insurance contexts (Joe, 2015).

Tail dependence, the tendency for extreme outcomes to occur simultaneously across variables, is a critical aspect of health insurance risk modeling. Studies by (Embrechts, Lindskog, and McNeil, 2003) highlighted the applicability of copulas in capturing such extremal dependencies. This is critical in health insurance, where the joint occurrence of large claims can have substantial financial implications.

Recent research has applied these models to health insurance claims data to better understand and predict the occurrence of large, infrequent claims. Zhang and Dukic, 2013 used copula models to investigate the dependencies between different types of health insurance claims, while Czado and Nagler, 2022 extended this work by employing vine copulas to capture the complex interactions within mixed data more effectively.

In terms of tail dependence, studies like those by Levantesi and Menzietti, 2017 have explored the use of copulas in long-term care insurance, where the occurrence of extreme events is of particular concern. Such models are crucial in predicting the longevity risk and pricing long-term care insurance products.

In summary, the literature indicates that tail-dependence copula-based regression models are an essential tool for understanding and managing the risks associated with health insurance data.

Copula-based regression models have garnered increasing attention in the statistical and machine-learning communities for their ability to capture complex dependencies between variables. Unlike traditional regression models that often assume independence or simple linear relationships among variables, copulas allow for a flexible representation of the dependence structure, accommodating tail dependencies and asymmetries often present in real-world data (Nelsen, 2006). This is particularly useful when dealing with mixed data types—comprising continuous, discrete, and categorical variables—where capturing the intricate relationships between variables is critical for accurate modeling and prediction (Genest and Favre, 2007).

In fields such as finance (Rodriguez, 2007), insurance, and environmental studies, where mixed data types are prevalent, the use of copula-based regression models can lead to more robust risk assessments and better-informed decision-making processes (Embrechts, McNeil, and Straumann, 2002). For instance, in health insurance, copula models can be employed to understand the relationship between patient demographics, past medical history, and the cost of insurance claims, thus enabling better risk pricing and policy design (Frees and Valdez, 1998).

The versatility of copulas stems from Sklar's theorem, which states that any multivariate joint distribution can be expressed as a copula function that connects the marginal distributions of individual variables (Sklar, 1959). This property is particularly advantageous for mixed data modeling, as it allows for separate modeling of the marginal distributions appropriate for each type of data, followed using a copula to model their dependence structure (Joe, 1997).

The current landscape of copula-based regression models for mixed data is rich, with various approaches tailored to different types of mixed data. For instance, vine copulas offer a flexible framework for high-dimensional mixed data modeling, decomposing a multivariate copula into a sequence of bivariate copulas, thus simplifying estimation and interpretation (Aas et al., 2009). Another approach involves the use of factor copulas, where latent factors are introduced to capture the underlying dependence among variables (Krupskii and Joe, 2015).

Despite their advantages, copula-based regression models are not without challenges. The selection of an appropriate copula function, estimation of copula parameters, and the handling of high-dimensional data remain active areas of research (Kurowicka and Joe, 2011).

In conclusion, copula-based regression models offer a powerful and flexible approach for analyzing and interpreting mixed data. As data complexity grows and the need for sophisticated modeling techniques becomes more pressing, copulas are likely to play an increasingly central role in statistical analysis and predictive modeling across various disciplines.

3. Methodology

In this study, we adopt copula-based regression models to analyze mixed health insurance data. We begin by specifying the appropriate marginal distributions for our data followed by the construction of joint distributions using copulas. The methodology emphasizes the use of copulas to capture the dependence structure between variables, particularly focusing on tail dependence characteristics.

3-1 Marginal Model Specification

For the continuous response variable representing health insurance charges, which is a continuous positive variable, a common choice for the marginal distribution of the response variable is the Gamma distribution due to its flexibility in modeling skewed positive data as:

$$f_Y(y;\alpha,\beta)=\frac{\beta^{\alpha}y^{\alpha-1}e^{-\beta y}}{\Gamma(\alpha)}\ ,y>0,\alpha>0,\beta>0 \tag{1}$$

Where y is the charge size of claims per month, α is the shape parameter, β is the rate parameter, and $\Gamma(\alpha)$is the gamma function evaluated at α

Also, For count variables such as the number of children, we specify the Negative Binomial distribution:

$$P(Y = y; r, p) = \binom{y + r - 1}{y} p^r (1 - p)^y, \; y = 0,1,2, \dots \tag{2}$$

Where y is the children's number, r is the number of failures until the experiment is stopped, and p is the probability of success in a single trial.

For binary explanatory variables like sex (male/female) and smoking (yes/no), we incorporate them into the model using a logistic regression framework.

3-2 Copula Types and Their Properties

3-2-1 Gaussian Copula

The Gaussian copula is derived from the multivariate normal distribution, capturing linear dependencies between variables. The strength and direction of the dependencies are measured by the correlation matrix, which does not capture tail dependence (Nelsen, 2006). The Gaussian copula is defined by its correlation matrix p and is given by:

$$C_{Gaussian}(u1, u2; \rho\,) = \Phi_\rho\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big) \tag{3}$$

Where $\Phi_\rho$is the cumulative distribution function (CDF) of the bivariate normal distribution with correlation coefficient ρ, and Φ^{-1} is the quantile function of the standard normal distribution.

3-2-2 Clayton Copula

The Clayton copula is known for its ability to model lower tail dependence, meaning it can effectively capture the scenario where extremely low values occur simultaneously across variables. It features an asymmetry that makes it suitable for modeling data where such joint extreme values are more likely on the lower end (Hofert and Scherer, 2011).

The Clayton copula is defined by its parameter $\theta > 0$

$$C_{Clayton}(u1, u2; \theta\,) = \max\{(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, 0\} \tag{4}$$

The Clayton copula has lower tail dependence and is asymmetric.

3-2-3 Gumbel Copula

Conversely, the Gumbel copula is adept at modeling upper tail dependence, which is pertinent for insurance data where the primary concern is the co-occurrence of large claims. It is asymmetric and places more weight on the joint tails of the distribution (Nelsen, 2006).

The Gumbel copula is defined by its parameter θ≥1 and is given by:

$$C_{Gumbel}(u1, u2; \theta\,) = \exp\left\{-\left[(-log u_1)^\theta + (-log u_2)^\theta\right]^{\frac{1}{\theta}}\right\} \tag{5}$$

The Gumbel copula exhibits upper tail dependence.

3-2-4 Frank Copula

The Frank copula can capture both lower and upper tail dependencies, although it does not exhibit tail dependence explicitly. It is often used for its flexibility in modeling different levels of dependency between variables (Durrleman et al., 2000).is defined by:

$$C_{Frank}(u1, u2; \theta\ ) = -\frac{1}{\theta}\log(1 + \frac{(\exp(-\theta\, u_1)-1)(\exp(-\theta\, u_2)-1)}{\exp\ (ex\,p(-\theta\ )-1}) \tag{6}$$

Where θ is a non-zero real number that governs the dependency level.

3-2-5 Student's t Copula

The Student’s t copula extends the Gaussian copula to include tail dependence. It is symmetric and can capture the tail dependence in both the lower and upper tails, making it particularly useful for modeling data where extreme values are correlated (Demarta and McNeil, 2005).

The Student's t copula, defined by its correlation coefficient ρ and degrees of freedom ν, is given by:

$$C_t(u1, u2; \rho, \nu\ ) = T_{\nu,\rho}(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)) \tag{7}$$

Where $T_{\nu,\rho}$ is the CDF of the bivariate Student's t-distribution with ν degrees of freedom and correlation ρ, and t_ν^{-1} is the quantile function of the univariate Student's t-distribution with ν degrees of freedom.

3-3 Joint Model Specification

To construct the joint distribution of the response variable and the covariates, we employ Sklar's theorem, which states that any multivariate cumulative distribution function can be expressed in terms of univariate marginal distribution functions and a copula that describes the dependence structure between the variables (Sklar, 1959) $Y_1, Y_2, \dots, Y_n$ with marginal CDFs H is:

$$H(y_1, y_2, \dots, y_n) = C(F_1(y_1), F_2(y_2), \dots, F_n(y_n)) \tag{8}$$

and the corresponding joint probability density function (PDF) is:

$$h(y_1, y_2, \dots, y_n) = c(F_1(y_1), F_2(y_2), \dots, F_n(y_n)) \tag{9}$$

For each copula, we estimate the parameters using the Inference Functions for Margins (IFM) method, which is a two-step procedure. First, we estimate the parameters of the marginal distributions, and then we estimate the copula parameters using the pseudo-observations derived from the marginal models (Joe, 1997).

4. Estimation methods:

Fitting a tail-dependence copula-based regression model commonly involves two main steps: the estimation of marginal distributions and the estimation of the copula parameters. Here, we'll focus on the use of parametric methods and the maximum likelihood estimation (MLE) method.

Step 1: Estimation of Marginal Distributions:

Suppose we have a dataset of n observations of a random vector X = (X1, X2, ..., Xd), and we want to fit a copula model to this data.

The first step is to transform each variable to be marginally Uniform (0,1). This can be done in several ways, including empirical distribution function transformation or parametric transformation, this is often done by fitting a parametric model to each margin and applying the cumulative distribution function (CDF) to each data point (Genest and Favre, 2007).

Step 2: Estimation of Copula Parameters:

Parameter estimation can be done using different methods like the Inference Functions for Margins (IFM) method, Maximum Likelihood Estimation (MLE), Pseudo Maximum Likelihood (PML)and Kendall's tau methods

- **IFM Method:**
  1. Estimate the marginal distributions of XX and YY separately.
  2. Transform the marginals to a uniform using the estimated CDFs: U=(X), V=FY(Y).
  3. Estimate the copula parameters $\theta\theta$ using the pseudo-observations UU and VV, (Joe, 2005).

- **MLE: can be defined as:**

$$\hat{\theta} = \arg\max \sum_{i=1}^{n} \log\ c(u_i, v_i, \theta) \tag{10}$$

Where (u, ;) is the density of the copula and (ui,) are the pseudo-observations.
A copula model can be fit to the data. Different copula models permit different amounts and types of dependence, including tail dependence, (Nelsen, 2006).

- **PML:**

This is like IFM, but it estimates the copula parameters using the empirical cumulative distribution function (CDF) of the data rather than the fitted CDF.

- **Kendall's tau:**

A non-parametric measure of correlation between two random variables (Ferrario et al., 2008). In the context of copulas, Kendall's tau can be used to estimate the parameter of a copula that captures the dependence structure between the variables.

Step 3: Copula-Based Regression Model

To integrate copulas into regression:

- Assume Y=(X)+ϵY=g(X)+ϵ, where $\epsilon\epsilon$ is a random error.
- Model (X,)(X,Y) using a copula, which links the marginal distribution of XX to the conditional distribution of YY given XX.

The regression function (X)g(X) can be estimated non-parametrically or parametrically, and the dependence between XX and $\epsilon\epsilon$ is modeled by the selected copula

Step 4: Model Evaluation

Various statistical criteria and tests can be used to assess the goodness-of-fit and compare different copula models. Among these methods, the Akaike Information Criterion (AIC) is widely used.

- **Akaike Information Criterion (AIC)**

The Akaike Information Criterion (AIC) is a well-established method used to compare different parametric models based on their fit to the data while penalizing for the number of parameters used (Burnham and Anderson, 2002)., and defined by:

$$\text{AIC}=2\text{k}-2\ln(\text{L}) \tag{11}$$

where:
k is the number of parameters in the model, and L is the maximum value of the likelihood function for the model.

- **Residual Analysis**: Check residuals for independence and appropriate distribution.
- **Predictive Performance**: Use cross-validation or out-of-sample testing to assess predictive accuracy.

This approach provides a structured way to analyze dependencies in the tails of distributions, which is critical for accurate risk assessment in fields such as finance and insurance.

5. Application to health insurance data

5-1 Data description:

In this section, we apply the proposed copula-based regression to the medical insurance claim data which were selected from ACME Insurance Inc. which offers affordable health insurance to thousands of customers all over the United States, using information sourced from Kaggle (https://www.kaggle.com/), contains 1338 observations that consists of the following seven risk factors such as their age, sex, BMI, children, smoking habits and region of residence.

Data variables:

1. Age: The age of the main person covered by the insurance.
2. Sex: The gender of the person buying the insurance, either female or male.
3. BMI: Body Mass Index, a measure that compares weight to height to assess if a person's weight is high or low for their height. It's calculated as weight in kilograms divided by height in meters squared. A healthy BMI is typically between 18.5 to 24.9.
4. Children: No. of children or dependents are covered by health insurance.
5. Smoker: Indicates if the person buying the insurance smokes.
6. Region: The area where the insured person lives in the US, such as the northeast, southeast, southwest, or northwest.
7. Charges: The medical costs that the health insurance bills to an individual.

We split the data set into two parts: the training set (70%) with 936 observations and the test set (30%) with 402 observations. The training set is used to fit the models, while the test set is used to test the model and evaluate the accuracy.

The data set was used for exploratory data analysis including the examination of summary statistics, variable distributions, and relationships between variables to inform the choice of predictive models and feature selection, as shown in Table (1)

Table 1. Descriptive statistics for data

	Age	BMI	children	charges
Min.	18.00	15.960	0	1121.874
Q1	27.000	026.296	0	4740.287
Median	39.000	30.663	1	9382.033
Mean	39.207	0.07485	1.095	13270.422
Q3	51.000	034.694	2	16639.913
Max	64.000	53.130	5	63770.428
StD	14.050	6.098	1.205	12110.011

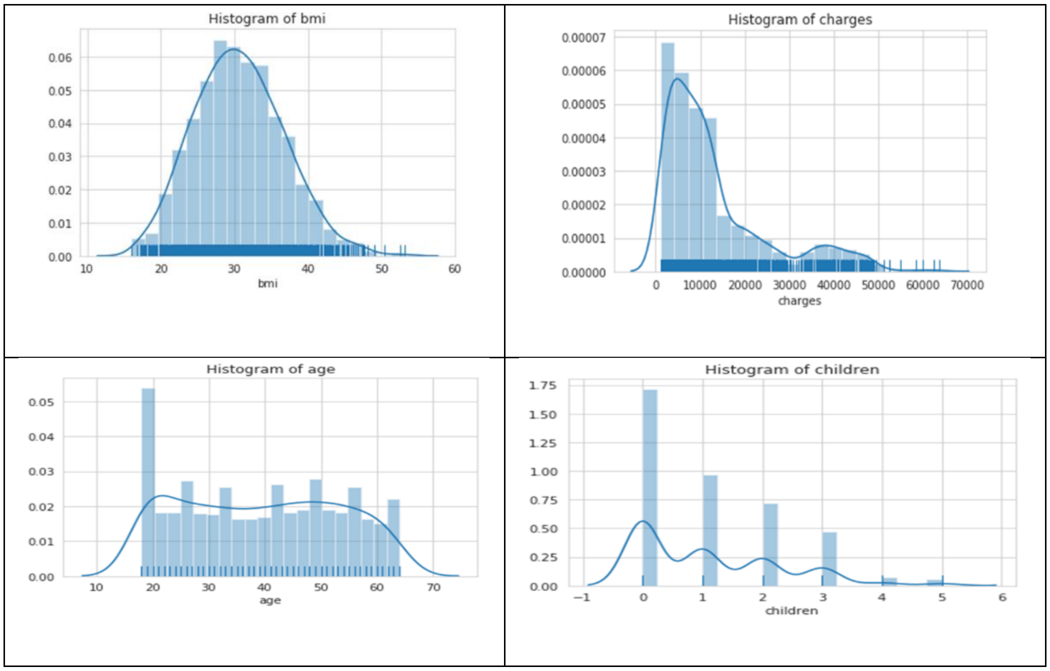
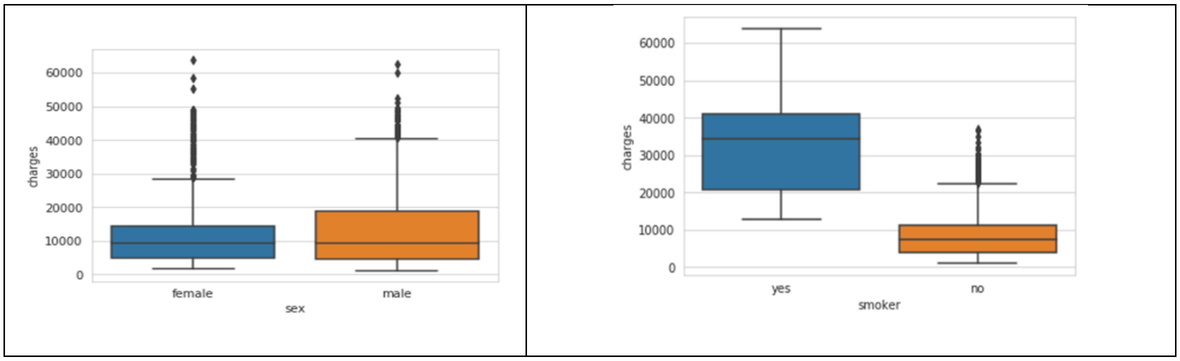


Figure 1. Histogram of numerical variables

From Figure (1) BMI data is the only feature that follows a normal distribution, with an average slightly above 30, which is higher than the typical maximum healthy value. The age data is mostly uniform but has more entries for younger ages. Also, the data for the number of children and medical charges are both skewed to the right, this skewness in the number of children is expected as most people nowadays tend to have fewer children or none, and older parents no longer count their grown children as dependents.

The skewness in medical charges shows that a few people have much higher charges than average, which could skew the results of the study. This skewness and the non-normal distribution of these features explain why there is a low correlation between them.



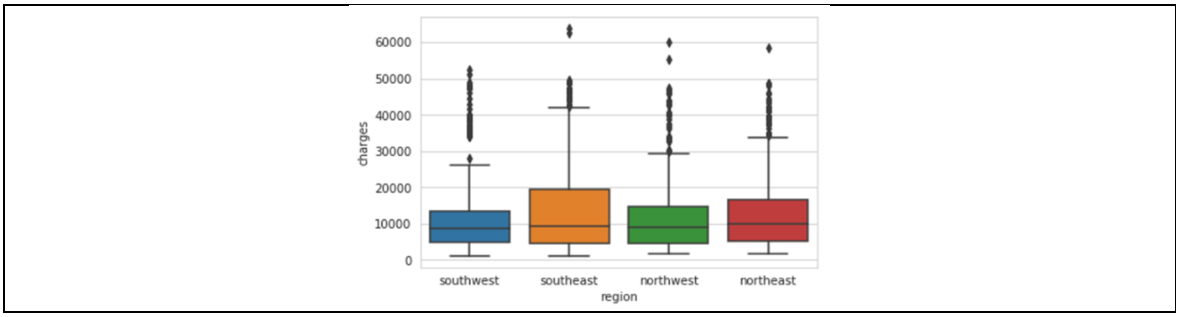


Figure 2. Boxplot of descriptive variables

The box plot analysis indicates significant differences in healthcare charges between males and females, with males incurring higher median charges and exhibiting greater variability. Both groups show outliers indicating occasional extremely high charges. Also, the box plot shows that individuals who smoke incur significantly higher healthcare charges compared to non-smokers. The median charges for smokers are notably higher and the range of charges (as represented by the box and whiskers) is broader, indicating greater variability in the charges they face. Non-smokers not only have lower median charges but also a more compressed range of charges, suggesting less variability. Finally, the box plot displays healthcare charges across four regions, all regions have outliers indicating extremely high charges, with the Southeast showing the most extremes.

This suggests that while there is variability in healthcare charges across all regions, the Southeast tends to be more expensive on average, with more frequent occurrences of very high charges.

Table 2. The correlation matrix between the features

	age	**BMI**	**children**	**charges**
age	1.000000	0.109272	0.042469	0.299008
BMI	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

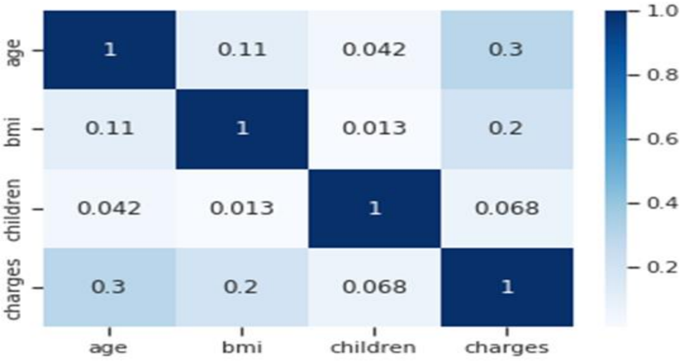


Figure 3. Correlation heatmap

Proceeding with a multivariate analysis, we will create a heatmap to illustrate the correlations among our data variables, as depicted in Figure (3). This visual representation through the heatmap will facilitate the examination of the correlations among different variables. We are particularly focused on uncovering the connections between charges and various demographic and medical data elements. In the heatmap, the intensity of the colors signifies the magnitude of the correlation between pairs of variables. Such a correlation matrix acts as a concise overview of our dataset, providing a basis for subsequent investigations, or serving as a diagnostic instrument for more complex analyses.

Before analysis, the dataset was preprocessed and cleaned. This process included checking for missing values, outliers, and inconsistencies. Missing values were imputed or removed, and outliers were removed based on the specific variable and its distribution.

Multicollinearity in multiple regression occurs when two or more predictors are closely related, meaning one predictor can predict another. This situation complicates estimating the individual impact of each predictor on the outcome.

One way to identify multicollinearity is by using the Variance Inflation Factor (VIF), which measures how much the variance of an estimated coefficient increases when multiple predictors are used compared to when the predictor is used alone. A VIF of 1 suggests no multicollinearity, while values between 5 and 10 usually indicate its presence. To resolve multicollinearity, you can remove the predictor with the highest VIF

Table 3. VIF for variable

	age	**BMI**	**children**	**smoker**	**region**
GVIF	1.016188	1.104197	1.003714	1.006369	1.098869
Df	1	1	1	1	3
GVIF^(1/(2*Df))	1.008061	1.050808	1.001855	1.003179	1.015838

None of the predictors in our case has a high value of VIF. Hence, we don't need to worry about multicollinearity in our case.

5-1 Modeling

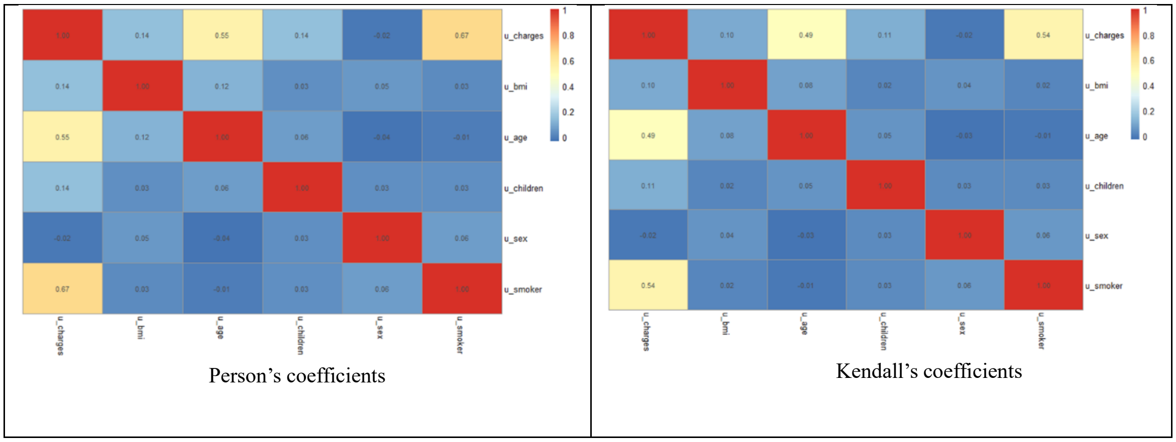
We used the "copula" package in R to find the best-fitted copula model for bivariate and multivariate cases. Table (4) below details the results of various copula methods and correlation measures and reveals significant insights into the level of binary dependence between insurance charges and several explanatory variables, specifically BMI, age, number of children, sex, and smoking status

Table 4. Level of binary dependence between charge and explanatory variables

Methods	**charges**					
	parameter	**BMI**	**age**	**children**	**sex**	**smoker**
Gaussian Copula	***rho***	0.1269	0.6972	0.1465	-0.02003	0.7481
	Std.	0.032	0.024	0.026	0.044	0.027
Clyton Copula	***alpha***	0.1763	1.93	0.2066	-0.02519	2.33
	Std.	0.049	0.162	0.041	0.055	0.242
Gumbel Copula	***alpha***	1.088	1.965	1.103	-	2.165
	Std.	0.024	0.081	0.021	-	0.121
Frank Copula	***alpha***	0.7332	5.577	0.8487	-0.1148	6.476
	Std.	0.005	0.168	0.004	0	0.278
t-Copula	***rho***	0.1269	0.6972	0.1465	-0.02003	0.7481
	Std.	0.032	0.024	0.026	0.044	0.027
Spearman's ρ	***rho***	0.116109	0.5526922	0.1200489	-0.01561484	0.6587308
	Std.	0.03039254	0.02550118	0.03037821	0.03059577	0.02302242
Kendall's τ	***tau***	0.08102956	0.4911357	0.0936326	-0.01275542	0.5381027
	Std.	0.03049888	0.02665471	0.03046507	0.03059701	0.02579169

We see from Table (4) that Both age and smoking status show strong dependence on insurance charges across most methods. This is evident from the high parameter values in Gaussian Copula, Clayton Copula, Gumbel Copula, Frank Copula, and t-Copula, as well as high correlation coefficients in Spearman's ρ and Kendall's τ. This suggests that age and smoking status are significant predictors of insurance charges. There is a moderate level of dependence between BMI and insurance charges, indicated by consistent moderate values across the Gaussian Copula, t-Copula, and other correlation measures. This suggests that BMI is a relevant but less potent predictor compared to age and smoking status. The dependence between the number of children and insurance charges is generally low to moderate, as reflected in the correlation coefficients and copula parameters. This indicates that while the number of children has some influence on insurance charges, it is less significant compared to other factors like age and smoking status. The parameter and correlation values associated with sex are very low and even negative in some cases, suggesting a very weak or negligible dependence between gender and insurance charges. This indicates that sex is not a significant predictor of insurance charges in this analysis.

The standard errors provided with each parameter estimate are relatively small, indicating a high level of precision in the estimates. This is crucial for confirming the reliability of the observed dependencies.



Person's coefficients

Kendall's coefficients

Figure 4. Correlation heatmap

The correlations matrix heatmap as shown in Figure (4) provides a clear depiction of how various factors are interrelated with insurance charges. Smoking status and age emerge as the most influential factors on insurance charges, underscoring their importance in risk assessment and pricing strategies in the insurance industry. Other factors like BMI, number of children, and sex show weaker associations, suggesting that their roles in insurance pricing are comparatively limited. This analysis is crucial for stakeholders in making informed decisions regarding policy formulation and risk management.

5-2-1 Bivariate Model

The scatterplot visualizes the relationship between Body Mass Index (BMI) and insurance charges using data points and a fitted line from an ordinary Generalized Linear Model (GLM).

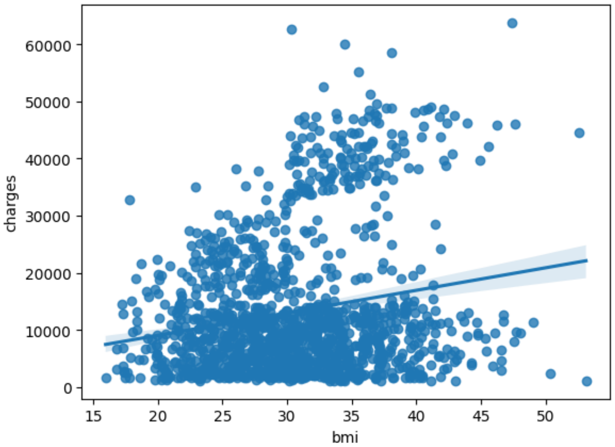


Figure 5. Binary GLM model

Figure (5) shows a positive correlation between BMI and charges which means that people with higher BMI, as they are at higher risk, spend more on medical insurance.

The study employs multiple modeling techniques, including ordinary Generalized Linear Models (GLM) and various copulas (Gaussian, Clayton, Frank, Gumbel, and t-copula), to explore the relationship between insurance charges (dependent variable) and Body Mass Index (BMI, explanatory variable). The models compare normal, gamma, and negative binomial (NB) families, focusing on estimates, standard errors, significance levels (p-values), and correlation coefficients (ρ)

Gamma and Negative Binomial distributions offer specific advantages in these contexts, particularly for modeling skewed and over-dispersed data (Cameron and Trivedi, 1998)

Table 5. Bivariate models

Model	family	Estimate		Std. Error		Pr(>|z|)		roh	Std. error
		Intercept	BMI	Intercept	BMI	Intercept	BMI		
GLM ordinary	normal	-621.55	452.99	2021.22	64.61	0.759	4.52e-12 ***		
	gamma	1.496e-04	-2.326e-06	1.049e-05	3.076e-07	< 2e-16 ***	9.36e-14 ***		
	NB	8.412204	0.034546	0.139042	0.004444	< 2e-16 ***	7.67e-15 ***		
Gaussian copula	normal	-967.25	464.25	1989.98	63.57	0.627	6.04e-13 ***	**0.1546**	**0.035**
	gamma	1.513e-04	-2.375e-06	1.030e-05	3.003e-07	< 2e-16 ***	7.42e-15 ***		
	NB	8.393335	0.035109	0.136923	0.004374	<2e-16 ***	1e-15 ***		
Clyton copula	normal	-700.78	455.43	1983.69	63.34	0.724	1.33e-12 ***	**0.2194 alpha**	**0.055**
	gamma	1.504e-04	-2.347e-06	1.034e-05	3.017e-07	< 2e-16 ***	1.91e-14 ***		
	NB	8.421309	0.034214	0.136516	0.004359	< 2e-16 ***	4.2e-15 ***		
Frank copula	normal	99.61	428.69	2001.67	63.83	0.96	3.25e-11 ***	**0.8967 alpha**	**0.006**
	gamma	1.459e-04	-2.207e-06	1.045e-05	3.065e-07	< 2e-16 ***	1.23e-12 ***		
	NB	8.449728	0.033304	0.137631	0.004389	< 2e-16 ***	3.25e-14 ***		
Gumble copula	normal	-300.64	442.84	1998.14	63.88	0.88	7.73e-12 ***	**1.11 alpha**	**0.028**
	gamma	1.483e-04	-2.288e-06	1.044e-05	3.061e-07	< 2e-16 ***	1.8e-13 ***		
	NB	8.439756	0.033688	0.137459	0.004395	< 2e-16 ***	1.78e-14 ***		
t- copula	normal	-293.98	442.20	2002.24	63.96	0.883	8.73e-12 ***	**0.1546**	**0.035**
	gamma	1.484e-04	-2.287e-06	1.047e-05	3.070e-07	< 2e-16 ***	2.14e-13 ***		
	NB	8.430442	0.033955	0.137706	0.004399	< 2e-16 ***	1.17e-14 ***		

GLM ordinary models as shown in table (5) significant effects of BMI on charges across all families (normal, gamma, NB), with p-values consistently less than 0.05, indicating strong evidence against the null hypothesis of no effect. Gaussian copula maintains a similar pattern to ordinary GLM but with a slightly lower correlation coefficient (ρ=0.627) indicating moderate dependency between charges and BMI. Clayton and Frank's copulas suggest higher dependencies (ρ values of 0.724 and 0.8967, respectively), with Clayton showing stronger dependency than Gaussian but weaker than Frank. These models indicate a robust effect of BMI on charges. Gumbel and t-copula show the highest ρ values (1.11 and 0.883, respectively), suggesting the strongest dependence among the models tested.

Model Fit and Complexity:

The AIC and log-likelihood values as shown in table (6) across models suggest that the Gaussian, Clayton, and t-copula models provide competitive fits, with the t-copula and Gumbel copula giving slightly better AIC in certain instances.

The negative binomial family consistently shows a better fit across different copulas compared to the normal and gamma families, indicating its appropriateness in handling overdispersion in the data.

Table 6. Goodness of fit measures for bivariate model

Model	family	The goodness of fit measures	
		AIC	2 x log-likelihood
GLM ordinary	normal	20247	-19530.4550
	gamma	19546	
	NB	19536	
Gaussian copula	normal	20243	-19526.1680
	gamma	19542	
	NB	19532	
Clyton copula	normal	20245	-19528.4620
	gamma	19543	
	NB	19534	
Frank copula	normal	20251	-19534.0630
	gamma	19550	
	NB	19540	
Gumble copula	normal	20248	-19531.8150
	gamma	19547	
	NB	19538	
t- copula	normal	20248	-19531.5460
	gamma	19547	
	NB	19538	

Figure (6) provides a comprehensive view of the data distribution and the relationship between Body Mass Index (BMI) and insurance charges.

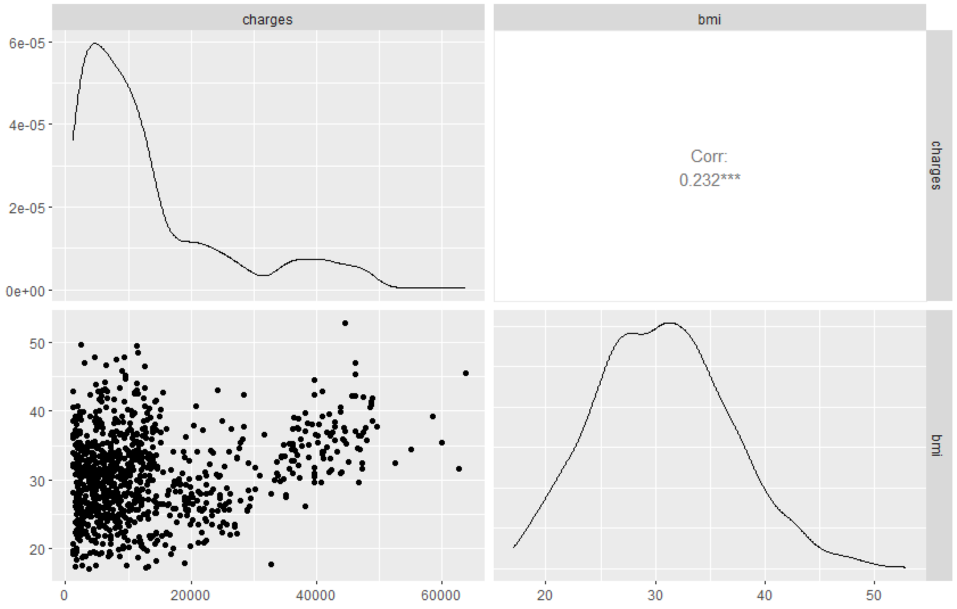


Figure 6. Relationship between (BMI) and insurance charges.

The combined analysis from the figure highlights a statistically significant but weak positive correlation between BMI and insurance charges. While there is a tendency for higher BMI to be associated with higher charges, the relationship is not strongly linear, and considerable variability in charges exists across different BMI levels. The density plots provide useful insights into the distribution patterns of both charges and BMI, showing a predominance of average BMI values and a higher frequency of lower insurance charges.

This visualization emphasizes the importance of considering other factors beyond BMI when predicting insurance charges, as the weak correlation suggests that BMI alone may not be a robust predictor of charge levels. Further analysis could explore additional variables that might influence insurance charges to develop a better understanding of the factor's insurance costs.

5-2-2 Multivariate Model

The presented results in Table (7) show the Estimated coefficients for various generalized linear models (GLMs) using different copulas (Gaussian, Clayton, t-copula, and Frank) and different family distributions (Normal, Gamma, Negative Binomial (NB)). The models evaluate the impact of several predictors—BMI, age, children, sex, and smoker status—on insurance charges The intercepts across models vary significantly, suggesting different baseline levels for the dependent variable when all predictors are at their reference levels.

BMI generally has a positive coefficient across models, indicating an increase in the dependent variable with higher BMI. The effect is statistically significant in most cases, Age also has a consistent positive effect across the models, showing that the dependent variable increases with age. The effect of having children on the dependent variable is mixed, with some models showing a positive effect and others showing no significant impact, The influence of variable sex is generally not significant or slightly negative in most models, suggesting minimal or adverse effects on the dependent variable. Finally, the status of being a smoker has a significant positive effect on the dependent variable across all models, indicating that smokers tend to have higher values of the dependent variable than non-smokers.

Correlation (Rho) and Standard Error indicate the model's estimation precision and the strength of relationships within the data structure, respectively. The significant and consistent impact of BMI, age, and smoker status across different models underscores their importance in predicting the dependent variable. The variability in the effects of children and sex suggests that these factors may interact with other variables or have context-specific impacts.

Table 7. Estimated coefficients for Multivariate models

Model	Family of GLM	coefficient	parameter						Rho	Std. error
			(Intercept)	Bmi	age	children	sex	smoker		
Gaussian copula	Normal	Estimate	-7261.07	315.49	252.72	219.59	-163.61	3329.71	0.1451	0.017
		Std. Error	1968.86	57.88	25.76	226.49	318.61	325.91		
		Pr(>|t|)	0.000239 ***	6.42e-08 ***	< 2e-16 ***	0.332537	0.607727	< 2e-16 ***		
	Gamma	Estimate	1.852e-04	-1.459e-06	-1.253e-06	-1.924e-06	1.400e-06	-1.378e-05		
		Std. Error	1.149e-05	2.969e-07	1.473e-07	1.224e-06	1.778e-06	1.382e-06		
		Pr(>|t|)	< 2e-16 ***	1.06e-06 ***	< 2e-16 ***	0.116	0.431	< 2e-16 ***		
	NB	Estimate	7.676512	0.025379	0.022019	0.026068	-0.015300	0.235785		
		Std. Error	0.134259	0.003947	0.001757	0.015445	0.021726	0.022224		
		Pr(>|z|)	< 2e-16 ***	1.27e-10 ***	< 2e-16 ***	0.0914 .	0.4813	< 2e-16 ***		
Clyton copula	Normal	Estimate	-7007.21	352.08	231.62	-96.49	-363.29	3192.95	0.3575 alpha	0.03
		Std. Error	2015.52	59.69	26.08	239.52	338.38	353.85		
		Pr(>(>|t|)	0.000531 ***	5.13e-09 ***	< 2e-16 ***	0.687148	0.283266	< 2e-16 ***		
	Gamma	Estimate	1.880e-04	-1.672e-06	-1.216e-06	-3.935e-07	1.125e-06	-1.381e-05		
		Std. Error	1.165e-05	3.078e-07	1.527e-07	1.242e-06	1.817e-06	1.555e-06		
		Pr(>(>|t|)	< 2e-16 ***	7.16e-08 ***	4.92e-15 ***	0.752	0.536	< 2e-16 ***		
	NB	Estimate	7.669944	0.030022	0.019959	-0.002187	-0.050205	0.252843		
		Std. Error	0.136663	0.004047	0.001768	0.016241	0.022943	0.023992		
		Pr(>|z|)	< 2e-16 ***	1.19e-13 ***	< 2e-16 ***	0.8929	0.0287 *	< 2e-16 ***		
t-copula	Normal	Estimate	-8342.39	365.10	236.95	107.85	175.59	3576.64	alpha 0.1451	0.017
		Std. Error	1994.62	58.59	25.61	234.64	324.72	342.99		
		Pr(>|z|)	3.16e-05 ***	6.97e-10 ***	< 2e-16 ***	0.646	0.589	< 2e-16 ***		
	Gamma	Estimate	1.875e-04	-1.543e-06	-1.211e-06	-2.233e-06	-8.431e-07	-1.308e-05		
		Std. Error	1.143e-05	3.152e-07	1.512e-07	1.246e-06	1.321e-06	1.261e-06		
		Pr(>|z|)	< 2e-16 ***	1.16e-06 ***	3.39e-15 ***	0.0735 .	0.5235	< 2e-16 ***		
	NB	Estimate	7.701521	0.026155	0.020416	0.027804	-0.0076	0.253076		
		Std. Error	0.136640	0.004013	0.001754	0.016073	0.022244	0.023496		
		Pr(>|z|)	< 2e-16 ***	17e-11 ***	< 2e-16 ***	0.0837 .	0.7322	< 2e-16 ***		
Frank copula	Normal	Estimate	-7724.57	330.10	256.63	285.38	-606.76	3943.55	1.102	0.004
		Std. Error	1965.53	57.99	25.34	226.57	323.78	337.13		
		Pr(>|z|)	9.12e-05 ***	1.68e-08 ***	< 2e-16 ***	0.2081	0.0612 .	< 2e-16 ***		
	NB	Estimate	7.686036	0.023387	0.023185	0.029300	-0.033385	0.286629		
		Std. Error	0.134045	0.003955	0.001728	0.015452	0.022081	0.022991		
		Pr(>|z|)	< 2e-16 ***	3.35e-09 ***	< 2e-16 ***	0.0579 .	0.1305	< 2e-16 ***		

Model Fit Statistics:

AIC and Log-Likelihood as shown in table (8) vary across models, with the Frank copula model generally showing a slightly better fit based on the AIC and log-likelihood values.

The Frank and Gaussian copula models generally provide a robust fit as indicated by lower AIC values and significant predictor effects.

Table 8. Goodness of fit measures for the multivariate model

Model	family	The goodness of fit measures	
		AIC	2 x log-likelihood
Gaussian copula	normal	20075	-19301.1350
	gamma	19366	
	NB	19315	
Clyton copula	normal	20100	-19317.6070
	gamma	19383	
	NB	19332	
Frank copula	normal	20047	-19267.7280
	gamma	-	
	NB	19282	
t- copula	normal	20066	-19300.65
	gamma	19365	
	NB	19315	

Figure (7) comprises a matrix of scatter plots, histograms, and correlation coefficients, providing a visualization of relationships and distributions among several variables: medical charges, BMI, age, number of children, sex, and smoking status.

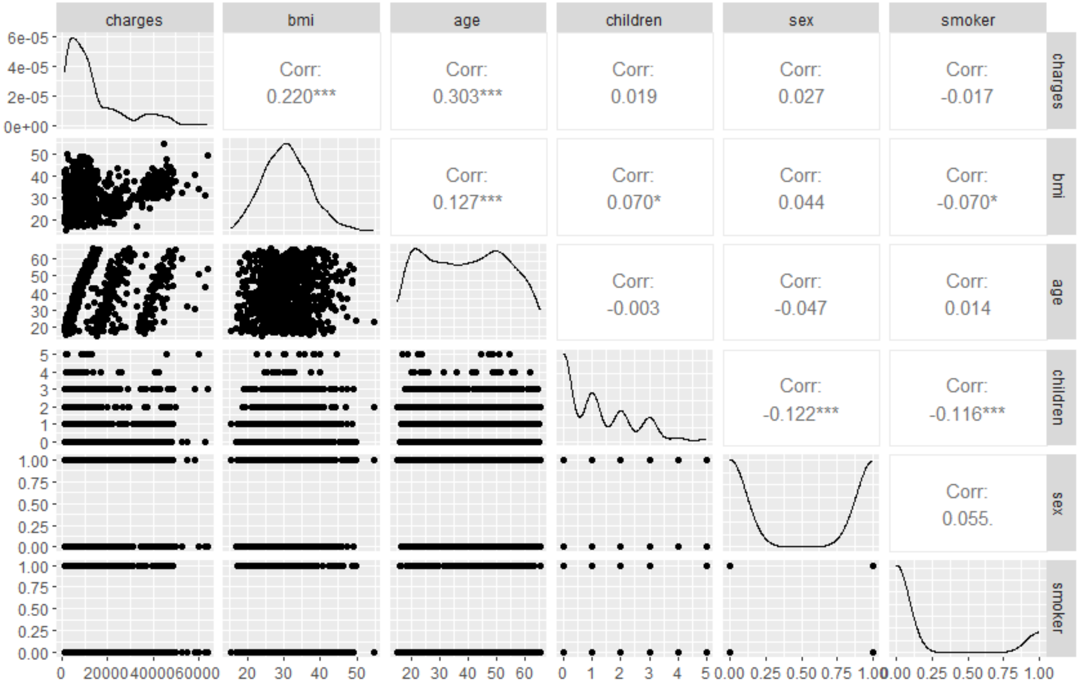


Figure 7. Matrix of scatter plots

This visualization aids in understanding the relationships between medical charges and various demographic factors. Age and BMI show positive correlations with medical charges, highlighting them as significant factors in predicting medical expenses. The correlations involving sex and smoker status with other variables suggest more complex relationships that might require further statistical or qualitative analysis to fully interpret. The data on children, sex, and smoker status provide insights into demographic and lifestyle patterns that could be pivotal in more detailed demographic studies or tailored health interventions.

6. Conclusion

In this study, when applying a bivariate model case, we found that BMI has a consistently significant impact on insurance charges across various models and families, reaffirming its relevance as an explanatory variable in insurance charge predictions. While ordinary GLM provides a strong baseline, copula models, especially the Frank and t-copula, demonstrate higher dependencies and slightly better model fits, suggesting their utility in capturing more complex relationships between the variables.

Insurance companies can benefit from using these sophisticated models to more accurately assess risk based on BMI, potentially leading to more tailored pricing strategies. Copula models may offer enhanced insights into the dependencies between charges and BMI, which could be crucial for risk segmentation and policy customization.

In the second case when applying multivariate models, the results of the study utilizing various copula models with Generalized Linear Models (GLMs) significantly demonstrate the intricate relationships between multiple variables such as BMI, age, number of children, sex, and smoking status in predicting the dependent variable. The application of Gaussian, Clayton, t-copula, and Frank copula models allowed for a nuanced understanding of these relationships, highlighting the importance of considering dependence structures in multivariate data analysis.

Significance of Smoking Status and BMI Across all models, the coefficients related to smoking status and BMI consistently showed significant p-values (almost all < 2e-16), indicating a strong influence on the dependent variable. This suggests that smoking status and BMI are critical factors in the studied context.

The Akaike Information Criterion (AIC) and the log-likelihood values vary across different copula families, indicating differences in model fit and efficiency in handling the data structure. In this study, the t-copula and Frank copula models generally provided a better fit (lower AIC and higher log-likelihood) compared to other models, suggesting their suitability in capturing tail dependencies and asymmetric relationships.

This research underscores the critical role of copula models in understanding and modeling dependencies among multiple variables in a multivariate setting. By employing different families of copulas:

The insights from this study are particularly valuable for practitioners and researchers involved in risk assessment, policy-making, and strategic planning where understanding complex, multivariate relationships is essential. By integrating the findings of this research, better-informed decisions can be made, tailored interventions can be designed, and more robust predictive models can be developed.

Declaration of interests
The authors declare that they have no conflict of interest.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2): 182-198.

Anderson-Cook, C. (2012). Quantitative risk management: concepts, techniques, and tools. *Journal of the American Statistical Association*, 101(476): 1731–1732

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media.

Cameron, A. C., and Trivedi, P. K. (1998). Regression Analysis of Count Data. Cambridge University Press.

Czado, C. (2010). Pair-Copula Constructions of Multivariate Copulas. In Copula Theory and Its Applications: 93-109. Springer, Berlin, Heidelberg.

Czado, C. and Nagler, T. (2022). Vine copula based modeling. *Review of Statistics and Its Applications*, (9): 453-477. doi.org/10.1146/annurev-statistics-040220-101153

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4): 278-305. 10.1080/03461238.2010.546147

Demarta, S. and McNeil, A. (2007). The t copula and related copulas. *International Statistical Review*, (73): 111-129

Durrleman, V., Nikeghbali, A.and Roncalli, T. (2000). A simple transformation of copulas. *SSRN Electronic Journal.* 1-15. doi.org/10.2139/ssrn.1032543

Embrechts, P., Lindskog, F., & McNeil, A. (2003). Modelling Dependence with Copulas and Applications to Risk Management. In Handbook of heavy tailed distributions in finance: 329-384. Elsevier.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependency in risk management: properties and pitfalls. In M. Dempster (Ed.), Risk management: value at risk and beyond, 176-223. Cambridge University Press.

Ferrario, A., Waters, H. R., & De Vries, C. G. (2008). A comparison of tail dependence estimates in claims reserving data. *Insurance: Mathematics and Economics*, 43(2): 255-262.

Frees, E. W., and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1): 1-25.

Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4): 347-368. doi.org/10.1061/(ASCE)1084-0699(2007)

Genest, C., Ghoudi, K., & Rivest, L. P. (2007). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3): 543-552.

Hofert, M. and Scherer, M. (2011). CDO pricing with nested Archimedean copulas. *Quantitative Finance*, 11(5): 775–787. 10.1080/14697680903508479

Joe, H. (1997). Multivariate Models and Dependence Concepts. Chapman & Hall.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2): 401-419. doi.org/10.1016/j.jmva.2004.06.003.

Joe, H. (2014). Dependence Modeling with Copulas. CRC press.

Kolev, N., & Paiva, D. (2009). Copula-based regression models: a survey. *Journal of Statistical Planning and Inference*, 139(11): 3847-3856.

Krupskii, A. and Joe, H. (2015). Factor copula models for item response data. *Psychometrika*, 80(1): 126–150. doi:10.1007/s11336-013-9387-4

Kurowicka,D. and Joe, H. (2010). Dependence Modeling: Vine Copula Handbook, ISBN: 978-981-4299-87-9

Levantesi, S. and Menzietti, M. (2017). Natural hedging in long- term care insurance, *Astin Bulletin* 48(1): 233-274. doi.org/10.1017/asb.2017.29

McNeil, A. J., Frey, R., & Embrechts, P. (2015). Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press.

Nelsen, R. B. (2006). An Introduction to Copulas. Springer Science & Business Media.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110: 4-18. doi.org/10.1016/j.jmva.2012.02.021

Rodriguez, J. C. (2007). Measuring financial contagion: A copula approach. *Journal of Empirical Finance*, 14(3): 401-423.doi.org/10.1016/j.jempfin.2006.07.002

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

Zhang, Y., and Dukic, V. (2013). Predicting multivariate insurance loss payments under the copula-based regression models. *The Journal of Risk and Insurance*, 2013, 80(4): 891–919. doi.org/10.1111/j.1539-6975.2012.01480.x