IARS

ORIGINAL ARTICLE

Regression Analysis Based on Copula Theory- By Using Gaussian family Copula

Mayooran Thevaraja[1] and Mezbahur Rahman[2]

[1]Department of Interdisciplinary Studies,
Faculty of Engineering, University of Jaffna,
Ariviyal Nagar, Killinochchi, Sri Lanka.
E-mail: mayooran@eng.jfn.ac.lk
[2]Department of Mathematics and Statistics
Minnesota State University, Mankato, MN 56001, USA
E-mail : mezbahur.rahman@mnsu.edu

Abstract

Researchers are often interested in studying the relationships between one variable and several other variables. Regression analysis is the statistical method for investigating such relationship, and it is one of the most commonly used statistical Methods in many scientific fields such as financial data analysis, medicine, biology, agriculture, economics, engineering, sociology, geology. However, the primary form of the regression analysis, ordinary least squares (OLS) is not suitable for actuarial applications because the relationships are often nonlinear, and the probability distribution of the response variable may be non-Gaussian distribution. One of the methods that have been successful in overcoming these challenges is the generalized linear model (GLM), which requires that the response variable have a distribution from the exponential family. In this research work, we study copula regression as an alternative method to OLS and GLM. The significant advantage of a copula regression is that there are no restrictions on the probability distributions that can be used. The first part of this study, we will briefly discuss copula regression by using several varieties of marginal copula functions and copula regression is the most appropriate method in a non-Gaussian variable (violated normality assumption) regression model fitting. Also, we validated our results by using real-world example data and random generated (50000 observations) data.

Keywords: Regression; Ordinary Least Squares (OLS); Multivariate Gaussian Copula; Copula Regression; Generalized Linear Models (GLM).

1. Introduction

The term "Regression" and the methods for investigating the relationships between two variables may date back to about hundred years ago. It was first introduced by Francis Galton in 1908, the renowned British biologist, when he was engaged in the study of heredity. One of his study was that the children of tall parents to be taller than average but not as tall as their parents. This "Regression toward mediocrity" gave this statistical methodology as their name. The term regression and its evolution primarily describe statistical relationship between the variables. The simple regression is the regression method to discuss the relationship between one response/dependent variable (y) and set of explanatory/independent variable (x). The basic ordinary least squares (OLS) regression model presents a specific model for the relationship. The distribution of Y given the co-variates assumed to be normal with a variance that is constant (that is, not related to the co-variates) and a mean that is related to the co-variates as $E(Y|X_1 = x_1,.., X_k = x_k) = \beta_0 + \beta_1 x_1 + \cdot \cdot + \beta_k x_k$. The equivalent (in this case) techniques of maximum likelihood and least-squares are used to estimate the unknown coefficients. To get more flexible approaches to apply in real world problems, different models have been proposed overcoming the less realistic assumptions of the Linear Models, where a Gaussian distribution was assumed for the dependent variable, with a constant variance and linear relationship between the predictor and the dependent variable. The development of the Generalized Linear Models (GLM) McCullagh and Nelder, 1989 relaxed the distributional assumption from Gaussian to any other distributions from the exponential family and the proposal of the