Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/130406

Bivariate Logit Models with Dummy Endogenous Regressors Using Copulas

Dawlah Alsulami[1,*], *Hadeel Klakattawi*[1], *Lamya Baharith*[1] *and Mervat Abd Elaal*[2,3]

[1]Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Statistics, Al-Azhar University, Cairo 11751, Egypt
[3]Canal High Institute of Engineering and Technology, Suez, Egypt

Received: 7 Dec. 2023, Revised: 21 Jan. 2024, Accepted: 19 Mar. 2024
Published online: 1 Jul. 2024

Abstract: Measuring the impact of a binary treatment on a binary response variable is of great interest in many medical, social and economic applications. Estimating such effect is very important when the endogeneity problem occurs. This research proposes a bivariate logit model to control endogeneity when the structural errors of the two equations are correlated. The copula approach will be applied to estimate the dependence between the binary treatment and the binary response; and hence, the joint normality assumption of the structural error is irrelevant. For estimation, the maximum likelihood method will be applied to estimate the model parameters. The performance of the copula bivariate logit model in estimating the dependence between the binary treatment variable and the binary response variable is assessed by the Average Treatment Effect (ATE) criterion in both simulation study and real medical data.

Keywords: Average treatment effect, Bivariate logit, Binary response, Copula, Maximum likelihood

1 Introduction

In many medical, social and economic applications, measuring the dependence between a binary treatment and a binary response variable is of vital importance. Therefore, the attention has been drawn; recently, to the importance of estimating such effect especially in the presence of endodgeneity, when the two binary variables depend on unobservable confounding variable which is hard to determine. The problem of endogeneity arises when a regression model lacks essential covariates, typically because they are not readily available, causing them to be included in the model's error term. This problem can't be controlled by the univariate probit and/or logit model as they will give biased estimation. Thus, many methods have been proposed to estimate the dependence between the binary treatment and response variable. One popular way to control endogeneity; especially in medical and economic studies, is to use the idea of instrumental variable which separates the effect of the predictors. Then, the estimation methods; such as, the generalized method of moment and the maximum likelihood (ML) can be applied.

Another useful method to control endogeneity is to apply the bivariate probit model. In this model two binary equations are used, and the errors of the two equations follow a bivariate Gaussian distribution with correlation parameter $\theta \neq 0$. Then, the model parameters can be estimated by the ML method. The bivariate probit model is widely applied in literature; for example, [1] applied this model on a patients with end-stage renal disease who have two choices. The first choice is dialysis modality and the other one is dialysis unit's type (private, public), which depend on unobserved factors; such as, patients' clinical factors and the characteristic of each unit. [2] used the recursive bivariate probit model to test the impact of diabetes on Canadian employment. [3] applied the recursive bivariate probit to estimate the relationship between the women's decision to participate in workface, and the formal hiring decision of organization. More applications for economic and health studies can be found in [4],[5], [6], [7], [8], [9] and [10]; among others. [11] proposed a flexible method to measure the impact of a binary treatment on a binary response when the endogeneity is present. They developed a two-stage generalized additive model for instrumental variable estimation and accounting for

* Corresponding author e-mail: dalsulami@kau.edu.sa

possible covariate nonlinear relationship by utilizing regression spline approach. They introduced a semiparametric recursive bivariate probit model to handle the endogeneity and the possible nonlinear effect of predictors. Their model is more effective than the classical bivariate probit model as it doesn't assume a specific relationship between the response and the continuous predictors, but rather applied the idea of penalized regression spline. [12] applied the semiparametric probit model to measure the effect of obesity on the employment probability in Italy. They assumed that both obesity and employment probability depend on unobserved confounding variable and hence the endogeneity problem occurs. [13] applied the two-stage generalized model to study the effect of Malawi women's education on fertility. [14] utilized a semiparametric bivariate probit model to analyse the dependence between the number of in-hospital deaths and the effectiveness outcome in ST-Elevation myocardial infraction patients. [15] proposed a simultaneous estimation method for the semiparametric recursive bivariate probit equation to deal with unobserved confounders, nonlinearity of the predictors and over dispersion.

These methods are effective in many situations; however, they assumed a Gaussian dependence between the errors of the response and treatment equations, which is violated in, may real world problems. Thus, [16] presented a copula bivariate probit model to account for the possible non-Gaussian dependence in the error terms. [17] introduced a new method to measure the impact of a binary treatment on a binary response. Their proposed model has the ability to control the effect of unobserved confounders, the possible nonlinear effect of the predictors and the possibility of non-Gaussian dependence between the error terms. [18] introduced a class of bivariate threshold crossing model, which includes the bivariate probit model as a special case, and used a parametric copula function to model the dependence between the error terms assuming that the marginal distribution of the errors is arbitrary but is known. [19] proposed an econometric model for estimating treatment effects in binary choice outcomes, employing a copula to capture the dependence of unobservable terms. The copula-based approach accommodates different dependence structures. Through a simulation study, he demonstrated that misspecifying the dependence structures leads to biased estimates of treatment effects. [20] introduced and implemented commands for estimating three distinct endogenous models of binary choice outcomes. All model estimations utilized copula-based maximum-likelihood estimation as the underlying statistical methodology. [21] investigated the link between financial inclusion and women's economic empowerment in Ethiopia using methods like endogenous switching regression and instrumental variables. Their results emphasize the positive impact of financial inclusion on women's economic empowerment. Although the probit model is widely used in literature, the logit model is more popular in many real applications because of its ability to interpret the coefficients in terms of odd ratio.

Thus to control for this form of unmeasured heterogeneity in the empirical context of this paper, we will utilize the bivariate logit model to estimate the effect of a binary endogenous treatment variable Y_1 on a binary response Y_2. The proposed model builds on a first equation modeling the endogenous dummy variable, a second equation is an outcome equation which determines the response variable Y_2 that depends on the endogenous binary regressor Y_1 and other covariates. The two equations are then connected via a bivariate logit distribution which makes it possible to model the correlation between the two equations, hence accounting for unobserved heterogeneity, assuming that the error terms follow bivariate logistic distribution.

The rest of the paper is organized as follows. In Section 2, we provide details on the model specification employed here. Subsection 2.1 discusses the ML method applied to estimate the model parameters. A simulation study will be presented in Section 3. To assess the performance of our model, an application to real medical data will be discussed in Section 4.

2 Bivariate Logit Model

The bivariate logit model offers a convenient method for gauging the impact of an endogenous binary regressor, denoted as y_1, on a binary outcome variable, denoted as y_2. The conventional model presupposes a consistent treatment effect, the existence of an exclusion restriction, and the absence of simultaneity. In a formal sense, the structural model encompasses the following pair of latent equations:

$$\begin{aligned} y_{1i}^{*} &= \boldsymbol{x}_{1i}'\boldsymbol{\alpha}_1 + u_{1i} \\ y_{2i}^{*} &= \beta y_{1i} + \boldsymbol{x}_{2i}'\boldsymbol{\alpha}_2 + u_{2i}, \qquad i = 1,2,...,n, \end{aligned} \tag{1}$$

where y_{1i}^* and y_{2i}^* are called latent variables for the binary variables y_{1i} and y_{2i}; respectively, such that:

$$y_{ji} = \begin{cases} 1 & \text{if} \quad y_{ji}^* > 0 \\ 0 & \text{if} \quad y_{ji}^* \leq 0, \end{cases} \qquad j = 1,2.$$

In Equation (1), the instrument variables $\boldsymbol{x}_{1i}' = (1, x_{11i}, x_{12i}, ..., x_{1pi})$ is the ith-row of an $(n \times p)$ matrix of regressors which affect y_{1i} but have no direct effect on the binary outcome variable y_{2i}. Similarly, $\boldsymbol{x}_{2i}' = (1, x_{21i}, x_{22i}, ..., x_{2qi})$ is the ith-row of an $(n \times q)$ matrix of regressors. Also, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are two $(p \times 1)$ and $(q \times 1)$ vectors of coefficients, respectively and β is the coefficient of the endogenous binary variable y_{1i}. Moreover, the error terms u_{1i} and u_{2i} are assumed to have the following joint distribution function:

$$F(u_{1i}, u_{2i}) = \Psi(u_{1i}, u_{2i}; \rho),$$

where $\Psi(.)$ is the Cumulative Distribution Function (CDF) of the bivariate logit distribution with coefficient of correlation $\rho \neq 0$. Therefore, to estimate the model parameters correctly, the dependence between the errors u_{1i} and u_{2i} should be taken into account. In this paper, a copula bivariate logit model is proposed to estimate the effect that a binary treatment variable has on a binary outcome variable, in the presence of endogeneity, by assuming different types of copulas and hence, different models will be generated. Therefore, the error terms u_{1i} and u_{2i} are assumed to have the following joint CDF:

$$F(u_{1i}, u_{2i}) = \mathscr{C}(F_1(u_{1i}), F_2(u_{2i}); \theta), \tag{2}$$

where, $\mathscr{C}(\cdot)$ is a copula function with dependence parameter $\theta \neq 0$ and both $F_1(u_{1i})$ and $F_2(u_{2i})$ are the univariate logistic CDF for u_{1i} and u_{2i}, thus

$$F_j(u_{ji}) = \frac{e^{u_{ji}}}{1 + e^{u_{ji}}}, \qquad j = 1,2. \tag{3}$$

It's crucial to understand that the suggested recursive bivariate logit model introduces two forms of dependence between y_{1i} and y_{2i}, linked to the parameters β and θ, respectively. Even though the joint model simplifies to two separate logit equations when the structural errors are independent $\theta = 0$, this doesn't imply independence between y_{1i} and y_{2i}. The second logit equation of the recursive base model determines the probability of y_{2i} conditional on y_{1i}, so complete independence requires $\theta = 0$ and $\beta = 0$. The copula bivariate logit model in this study employs copulas to depict dependence between the structural errors. It doesn't directly model the dependence between the two binary outcomes, but the dependence among the structural errors evidently influences that dependence. In this paper four different types of copula functions will be applied to estimate the effect of the treatment variable y_{1i} on the outcome variable y_{2i} in Model (1). The chosen copulas are: the Gaussian, Frank, FGM and Plackett copula. These copula functions are widely applied as they allow for negative dependence. Moreover, they are symmetric in both tails except for plackett copula. Table 1 defines the chosen copula with the parameter's domain.

Table 1: The Gaussian, Frank, FGM and Plackett copula function with parameter's domain

Type of Copula	$\mathscr{C}(u, v; \theta)$	Domain of θ
Gaussian	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta)$	$\theta \in (-1, 1)$
Frank	$-\frac{1}{\theta} log(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)})$	$\theta \in \mathbb{R} \{0\}$
FGM	$uv(1 + \theta(1-u)(1-v))$	$\theta \in (-1, 1)$
Plackett	$\frac{1}{2\theta}(1 + \theta(u+v)) - [1 + \theta(u+v)^2 - 4\theta(\theta+1)uv]^{\frac{1}{2}}$	$\theta \in [-1, \infty)$

2.1 Estimation strategy

2.1.1 Maximum likelihood

In the bivariate logit model, The primary focus is on determining the structural treatment parameter β, commonly known as the average treatment effect:

$E_x[P(u_{2i} > -\beta - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2) - P(u_{2i} > -\boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2)]$.
Let $\mathbf{x}$ be a vector of x_{1i} and x_{2i}, the joint distribution of y_{1i} and y_{2i} (conditional on x_{1i} and x_{2i}) has four elements:

$$\begin{aligned}
P(y_{1i}=0,y_{2i}=0|\mathbf{x}) &= P(u_{1i} \le -\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1, u_{2i} \le -\boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2),\\
P(y_{1i}=1,y_{2i}=0|\mathbf{x}) &= P(u_{1i} \le -\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1, u_{2i} \le -\beta - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2),\\
P(y_{1i}=0,y_{2i}=1|\mathbf{x}) &= P(u_{1i} \le -\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1, u_{2i} \le -\boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2),\\
P(y_{1i}=1,y_{2i}=1|\mathbf{x}) &= P(u_{1i} \le -\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1, u_{2i} \le -\beta - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2). \qquad (4)
\end{aligned}$$

This distribution is fully determined once the joint distribution of u_{1i} and u_{2i} is known. Thus from (2) and (3), the joint probability density function of y_{1i} and y_{2i} can be written compactly as:

$$f(y_{1i},y_{2i}|\mathbf{x}) = \mathscr{C}(F_1(\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(\beta y_{1i} + \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta).$$

Thus, under copula representation with logit marginal, the probability expressions in (4) can be written as:

$$\begin{aligned}
P(y_{1i}=0,y_{2i}=0|\mathbf{x}) &= \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(-\boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta],\\
P(y_{1i}=1,y_{2i}=0|\mathbf{x}) &= \mathscr{C}[1, F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta] - \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta],\\
P(y_{1i}=0,y_{2i}=1|\mathbf{x}) &= \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), 1; \theta] - \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(-\boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta],\\
P(y_{1i}=1,y_{2i}=1|\mathbf{x}) &= 1 - \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), 1; \theta] - \mathscr{C}[1, F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta],\\
&\quad + \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta].
\end{aligned}$$

The copula bivariate logit model's joint probabilities are influenced by both the chosen copula and four parameters, $\xi = (\beta, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \theta)$, where θ is the copula function's dependence parameter. If we assume the true copula belongs to a parametric family, a reliable and asymptotically normally distributed estimator for the parameter ξ can be derived through ML. The likelihood function can be formulated as:

The likelihood function an be expressed as:

$$\begin{aligned}
L(\xi) &= \prod_{i=1}^{n} P(y_{1i}=1,y_{2i}=1)^{y_{1i},y_{2i}} \times P(y_{1i}=1,y_{2i}=0)^{y_{1i},(1-y_{2i})}\\
&\quad \times P(y_{1i}=0,y_{2i}=1)^{(1-y_{1i}),y_{2i}} \times P(y_{1i}=0,y_{2i}=0)^{(1-y_{1i}),(1-y_{2i})}. \qquad (5)
\end{aligned}$$

Numeric optimization techniques can be applied to maximize the log-likelihood function in (5). A crucial condition for identification is the presence of at least one exogenous regressor with a non-zero coefficient, denoted by $\boldsymbol{\alpha}_1 \neq 0$, or $\boldsymbol{\alpha}_2 \neq 0$. Assuming the model is accurately specified, the ML estimator exhibits standard asymptotic properties. These estimators are beneficial as they offer optimal approximations to an undisclosed true model.

2.1.2 The average treatment effect

The effect of a binary treatment y_{1i} on a binary outcome $y_{2i} = 1$ is of vast interest in many real applications. One of the most popular measures used in literature is the ATE. This measure compares the expected value of the outcome with and without the treatment. The Sample Average Treatment Effect (SATE) is calculated by:

$$SATE(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} P(y_{2i}=1|y_{1i}=1,\mathbf{x}) - P(y_{2i}=1|y_{1i}=0,\mathbf{x}),$$

where,

$$\begin{aligned}
P(y_{1i}=1,y_{2i}=1|\mathbf{x}) &= 1 - \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), 1; \theta] - \mathscr{C}[1, F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta]\\
&\quad + \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta],\\
P(y_{1i}=1,y_{2i}=0|\mathbf{x}) &= \mathscr{C}[1, F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta] - \mathscr{C}[F_1(-\boldsymbol{x}'_{1i}\boldsymbol{\alpha}_1), F_2(-\beta y_{1i} - \boldsymbol{x}'_{2i}\boldsymbol{\alpha}_2); \theta].
\end{aligned}$$

3 Simulation Study

In order to assess the performance of the proposed methodology, we generate data from the following model:

$$y_1 = 1(\alpha_1 x_1 + u_1 > 0), \qquad (6)$$

$$y_2 = 1(\alpha_0 + \beta y_1 + \alpha_2 x_2 + u_2 > 0), \qquad (7)$$

where the coefficients are sitting to $\alpha_1 = 0.4, \alpha_0 = 0.9, \alpha_2 = -0.5$ and $\beta = 0.8$ and the instrumental variable x_1 and the regressor x_2 are two independent and identically distributed random variables with mean 0 and variance 1. The error terms u_1 and u_2 are drawn from different types of copula with dependence parameter $\theta \neq 0$. For Model (7), we will consider the four copula functions; Gaussian, FGM, Frank and Plackett with different sample sizes.

The main interest here is to investigate the effect of the treatment y_1 on the probability that the response variable $y_2 = 1$. This effect can be measured by the idea of the Sample Average Treatment Effect (SATE), which is defines by the difference between the expected value of the response variable with and without the presence of the treatment y_1. Therefore:

$$SATE = \frac{1}{n}\sum_{i=1}^{n} \Psi(0.9+0.8-0.5x_1) - \Psi(0.9-0.5x_2),$$

where $\Psi(.)$ denotes the conditional distribution function of the logistic distribution. To compare between the four alternative copulas, we calculate the Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{\mid SATE - TrueATE \mid}{\mid TrueATE \mid} \times 100,$$

where the true ATE is approximated by averaging $\Psi(0.9+0.8-0.5x_1) - \Psi(0.9-0.5x_2)$ for the sample size of 5 million and equals to 0.1336622. The dependence parameters for the four copulas will be chosen to give the same value of the Kendall's Tau as the direct comparison between their dependence parameters can't be made. Hence, the dependence parameters for the four copulas are chosen to be 0.23,1.32, 0.675 and 2.08; respectively, which give the value of $\tau = 0.15$ for the kendall's Tau.

Tables 2-5 show the simulation results for the four copulas; Gaussian, FGM, Frank and Plackett with the dependence parameters 0.23, 1.32, 0.675 and 2.08; respectively, with sample sizes $N = 1000, 5000$, 10000 and 20000 and number of replications = 250.

Table 2: Simulation results for the parameter estimates and corresponding MSEs (in parentheses) for Gaussian copula DGP with $\theta = 0.23$

	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\beta}$	$\hat{\beta}_2$	*ATE*	*MAPE*	$-log\ likelihood$
			N=1000				
Gaussian	0.40506	0.88531	0.78088	-0.48303	0.12711	4.90101	-1163.090
	(0.00473)	(0.10540)	(0.57333)	(0.00642)			
FGM	0.40794	0.83502	0.93470	-0.49124	0.15373	15.02123	-1163.203
	(0.00466)	(0.08060)	(0.36056)	(0.00613)			
Frank	0.40437	0.89244	0.75434	-0.48081	0.12222	8.55794	-1163.106
	(0.00480)	(0.11775)	(0.65046)	(0.00666)			
Plackett	0.40397	0.89377	0.74697	-0.48027	0.12043	9.89420	-1163.122
	(0.00486)	(0.11568)	(0.66953)	(0.00680)			
			N=5000				
Gaussian	0.39998	0.88770	0.81872	-0.50126	0.13586	1.64522	-5825.337
	(0.00101)	(0.02648)	(0.14181)	(0.00132)			
FGM	0.40032	0.87860	0.86338	-0.50348	0.14267	6.74599	-5825.509
	(0.00100)	(0.02328)	(0.11765)	(0.00131)			
Frank	0.39996	0.89543	0.82053	-0.50172	0.13558	1.43931	-58325.474
	(0.00101)	(0.02737)	(0.14484)	(0.00134)			
Plackett	0.40001	0.89543	0.82119	-0.50170	0.13568	1.51380	-5832.482
	(0.00101)	(0.02737)	(0.14447)	(0.00132)			
			N=10000				
Gaussian	0.39712	0.88871	0.82078	-0.49717	0.13673	2.30218	-11662.561
	(0.00044)	(0.01070)	(0.05899)	(0.00070)			
FGM	0.39717	0.88843	0.84540	-0.49858	0.14004	4.77485	-11662.760
	(0.00044)	(0.01070)	(0.05841)	(0.00069)			
Frank	0.39709	0.89495	0.82970	-0.49790	0.13742	2.81522	-11662.788
	(0.00044)	(0.01213)	(0.06794)	(0.00071)			
Plackett	0.39716	0.89262	0.83594	-0.49805	0.13841	3.55469	-11662.832
	(0.00044)	(0.01193)	(0.06698)	(0.00071)			

Continued on next page

	N=20000						
	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\beta}$	$\hat{\beta}_2$	ATE	$MAPE$	$-log\ likelihood$
Gaussian	0.39996	0.89765	0.80463	-0.49870	0.13321	0.36693	-23313.94
	(0.00024)	(0.00539)	(0.03132)	(0.00039)			
FGM	0.40000	0.89724	0.83109	-0.50020	0.13674	3.00555	-23314.417
	(0.00024)	(0.00542)	(0.03203)	(0.00038)			
Frank	0.39993	0.90553	0.81194	-0.49952	0.13374	0.67830	-23314.397
	(0.00024)	(0.00624)	(0.03571)	(0.00039)			
Plackett	0.39998	0.90461	0.81472	-0.49955	0.13431	1.00252	-23314.443
	(0.00024)	(0.00624)	(0.03553)	(0.00039)			

Table 3: Simulation results for the parameter estimates and corresponding MSEs (in parentheses) for Frank copula DGP with $\theta = 1.32$

	N=1000						
	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\beta}$	$\hat{\alpha}_2$	ATE	$MAPE$	$-log\ likelihood$
Gaussian	0.40282	0.84515	0.86562	-0.49703	0.14115	5.60820	-1163.942
	(0.00436)	(0.10868)	(0.58245)	(0.00802)			
FGM	0.40510	0.80480	0.99287	-0.50474	0.16336	22.21908	-1164.000
	(0.00431)	(0.08706)	(0.39446)	(0.00758)			
Frank	0.40219	0.85949	0.82617	-0.49519	0.13458	0.69257	-1163.898
	(0.00437)	(0.11881)	(0.63173)	(0.00793)			
Plackett	0.40185	0.86162	0.81816	-0.49459	0.13306	0.44561	-1163.907
	(0.00444)	(0.11810)	(0.64116)	(0.00794)			
	N=5000						
Gaussian	0.39977	0.88256	0.80717	-0.49625	0.13483	0.88073	-5837.434
	(0.00085)	(0.01917)	(0.10727)	(0.00152)			
FGM	0.39973	0.89461	0.80310	-0.49784	0.13374	0.06413	-5837.299
	(0.00084)	(0.01581)	(0.08489)	(0.00148)			
Frank	0.39933	0.91079	0.76123	-0.49602	0.12672	5.19237	-5837.270
	(0.00085)	(0.02001)	(0.11632)	(0.00153)			
Plackett	0.39931	0.91027	0.76257	-0.49595	0.12688	5.07298	-5837.285
	(0.00085)	(0.02051)	(0.12041)	(0.00154)			
	N=10000						
Gaussian	0.40033	0.86235	0.85686	-0.49766	0.14319	7.13554	-11676.656
	(0.00044)	(0.01295)	(0.06834)	(0.00060)			
FGM	0.40014	0.87622	0.84823	-0.49882	0.14119	5.63693	-11676.472
	(0.00044)	(0.01128)	(0.05993)	(0.00058)			
Frank	0.39997	0.88719	0.82143	-0.49781	0.13675	2.31755	-11676.440
	(0.00044)	(0.01291)	(0.07192)	(0.00060)			
Plackett	0.39998	0.88708	0.82208	-0.49778	0.13684	2.38021	-11676.464
	(0.00044)	(0.01307)	(0.07295)	(0.00060)			
	N=20000						
Gaussian	0.40069	0.86172	0.86414	-0.49895	0.14456	8.16022	-23353.113
	(0.00021)	(0.00610)	(0.02911)	(0.00033)			
FGM	0.40037	0.87903	0.84869	-0.49993	0.14133	5.73883	-23352.65
	(0.00021)	(0.00513)	(0.02695)	(0.00033)			
Frank	0.40024	0.89042	0.82234	-0.49912	0.13704	2.53158	-23352.565
	(0.00021)	(0.00542)	(0.02912)	(0.00033)			
Plackett	0.40023	0.89159	0.81996	-0.49902	0.13664	2.23038	-23352.572
	(0.00021)	(0.00551)	(0.02983)	(0.00033)			

In Table 2, the Data Generating Process (DGP) is setting to Gaussian copula with $\theta = 0.23$. The Gaussian bivariate logit model and Frank bivariate logit model perform well in estimating the true ATE for all sample sizes. All models perform well whith $N = 20000$ with minimim MAPE equals to 0.36% produced by Gaussian copula. When the Frank

copula is used as a DGP with $\theta = 1.32$, Table 3 shows that Frank bivariate logit model model is the best models in estimating the true ATE with MAPE equals to 0.06% with $N = 5000$. However, it is less attractive with $N = 10000$ and $N = 20000$. In general, both Frank and Plackett perform well in estimating the true ATE. In Table 4, the DGP is sitting to

Table 4: Simulation results for the parameter estimates and corresponding MSEs (in parentheses) for FGM copula DGP with $\theta = 0.657$

	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\beta}$	$\hat{\alpha}_2$	ATE	$MAPE$	$-log\ likelihood$
N=1000							
Gaussian	0.39771	0.83130	0.86341	-0.48797	0.14102	5.50500	-1165.412
	(0.00451)	(0.13304)	(0.69206)	(0.00703)			
FGM	0.40078	0.78844	1.00473	-0.49769	0.16636	24.46379	-1165.517
	(0.00454)	(0.10021)	(0.41550)	(0.00640)			
Frank	0.39614	0.85221	0.80005	-0.48482	0.13039	2.44606	-1165.359
	(0.00464)	(0.14634)	(0.76467)	(0.00716)			
Plackett	0.39631	0.85005	0.80477	-0.48486	0.13121	1.83366	-1165.387
	(0.00466)	(0.14605)	(0.76348)	(0.00714)			
N=5000							
Gaussian	0.40132	0.85722	0.86872	-0.50077	0.14447	8.09162	-5832.249
	(0.00098)	(0.02486)	(0.13645)	(0.00131)			
FGM	0.40131	0.86972	0.86221	-0.50220	0.14303	7.01166	-5832.149
	(0.00098)	(0.02150)	(0.11338)	(0.00128)			
Frank	0.40090	0.88095	0.83181	-0.50064	0.13786	3.14302	-5832.153
	(0.00098)	(0.02586)	(0.14573)	(0.00132)			
Plackett	0.40096	0.87736	0.84109	-0.50087	0.13935	4.25835	-5832.190
	(0.00098)	(0.02573)	(0.14456)	(0.00133)			
N=10000							
Gaussian	0.39943	0.87880	0.82343	-0.50010	0.13765	2.98522	-11668.394
	(0.00050)	(0.01216)	(0.06535)	(0.00076)			
FGM	0.39916	0.89683	0.80577	-0.50105	0.13420	0.40326	-11668.183
	(0.00050)	(0.01036)	(0.05484)	(0.00074)			
Frank	0.39900	0.90488	0.78603	-0.50018	0.13091	2.05285	-11668.235
	(0.00050)	(0.01227)	(0.06717)	(0.00076)			
Plackett	0.39907	0.90128	0.79553	-0.50043	0.13244	0.90857	-11668.311
	(0.00050)	(0.01203)	(0.06531)	(0.00076)			
N=20000							
Gaussian	0.39976	0.88125	0.81855	-0.49837	0.13715	2.61410	-23341.696
	(0.00020)	(0.00569)	(0.02979)	(0.00033)			
FGM	0.39939	0.90091	0.79755	-0.49915	0.13305	0.45751	-23341.237
	(0.00020)	(0.00530)	(0.02884)	(0.00034)			
Frank	0.39932	0.90717	0.78327	-0.49858	0.13068	2.22759	-23341.351
	(0.00020)	(0.00618)	(0.03434)	(0.00034)			
Plackett	0.39938	0.90382	0.79200	-0.49879	0.13207	1.18988	-23341.485
	(0.00020)	(0.00613)	(0.03404)	(0.00033)			

FGM copula with $\theta = 0.67$. All models perform well in estimating the true ATE with the smallest MAPE equals to 0.4% produced by the FGM bivariate logit model with $N = 10000$. Moreover, both FGM and Plackett bivariate logit models outperform other models in estimating the true ATE for $N = 10000$ and $N = 20000$.

Table 5: Simulation results for the parameter estimates and corresponding MSEs (in parentheses) for Plackett copula DGP with $\theta = 2.08$

	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\beta}$	$\hat{\alpha}_2$	ATE	$MAPE$	$-log\ likelihood$
			N=1000				
Gaussian	0.39567	0.85609	0.82985	-0.49809	0.13543	1.32475	-1161.855
	(0.00452)	(0.11731)	(0.62155)	(0.00674)			
FGM	0.39825	0.80528	0.99546	-0.50913	0.16426	22.89743	-1161.971
	(0.00442)	(0.08957)	(0.37069)	(0.00659)			
Frank	0.39501	0.87727	0.77657	-0.49617	0.12640	5.43269	-1161.794
	(0.00457)	(0.12259)	(0.65992)	(0.00668)			
Plackett	0.39443	0.88350	0.75349	-0.49465	0.12205	8.68701	-1161.787
	(0.00463)	(0.12431)	(0.70165)	(0.00682)			
			N=5000				
Gaussian	0.39888	0.85477	0.86561	-0.49765	0.14441	8.04548	-5824.931
	(0.00076)	(0.02143)	(0.11533)	(0.00118)			
FGM	0.39898	0.85657	0.88946	-0.50029	0.14772	10.52464	-5824.827
	(0.00075)	(0.01798)	(0.09218)	(0.00115)			
Frank	0.39836	0.88281	0.82139	-0.49746	0.13643	2.07205	-5824.717
	(0.00076)	(0.02232)	(0.12901)	(0.00120)			
Plackett	0.39826	0.88673	0.81165	-0.49702	0.13478	0.84252	-5824.702
	(0.00076)	(0.02309)	(0.13556)	(0.00121)			
			N=10000				
Gaussian	0.39888	0.85856	0.85417	-0.49943	0.14310	7.06633	-11657.856
	(0.00045)	(0.01249)	(0.06484)	(0.00062)			
FGM	0.39878	0.86581	0.86572	-0.50155	0.14423	7.90975	-11657.659
	(0.00044)	(0.00991)	(0.05308)	(0.00062)			
Frank	0.39838	0.89027	0.80515	-0.49936	0.13434	0.51234	-11657.465
	(0.00045)	(0.01173)	(0.06835)	(0.00062)			
Plackett	0.39831	0.89506	0.79375	-0.49891	0.13248	0.87914	-11657.427
	(0.00045)	(0.01206)	(0.07067)	(0.00061)			
			N=20000				
Gaussian	0.40154	0.86804	0.84302	-0.49853	0.14125	5.68007	-11657.856
	(0.00022)	(0.00655)	(0.03466)	(0.00044)			
FGM	0.40130	0.87772	0.84891	-0.50027	0.14140	5.78934	-11657.659
	(0.00022)	(0.00522)	(0.02995)	(0.00042)			
Frank	0.40104	0.89761	0.80118	-0.49868	0.13366	0.00359	-11657.465
	(0.00022)	(0.00587)	(0.03493)	(0.00044)			
Plackett	0.40101	0.90071	0.79421	-0.49841	0.13252	0.84917	-11657.427
	(0.00022)	(0.00602)	(0.03575)	(0.00044)			

Table 5 summarizes the results with Plackett DGP and dependence parameter $\theta = 2.08$. In this case, both Frank and Plackett bivariate logit models work better than other models in estimating the true ATE with minimum MAPE equals to 0.003 obtained by the Frank bivariate logit model with $N = 20000$.

4 Real Data Application

To examine the effectiveness of the proposed methodology, we utilized the meps dataset available on R package GJRM and also can be obtained from[1]. These dataset include some information on personal health such as, private health cover, number of visiting doctors and health status. For the selection of variables in the proposed model, we followed the work

[1] http://www.meps.ahrq.gov/

of [22] and thus the bivariate logit model is written as:

$$\begin{aligned}\text{private} = {} & \text{bmi} + \text{income} + \text{age} + \text{education} + \text{as.factor(health)} + \text{as.factor(race)} \\ & + \text{as.factor(limitation)} + \text{as.factor(region)} + \text{gender} + \text{hypertension} \\ & + \text{hyperlipidemia} + \text{diabetes}\end{aligned}$$

$$\begin{aligned}\text{visits hospital} = {} & \text{private} + \text{bmi} + \text{income} + \text{age} + \text{education} + \text{as.factor(health)} \\ & + \text{as.factor(race)} + \text{as.factor(limitation)} + \text{as.factor(region)} + \text{gender} \\ & + \text{hypertension} + \text{hyperlipidemia} + \text{diabetes}\end{aligned}$$

To understand the variables used in the above model, a short descriptions of each variable is provided in Table 6. To

Table 6: Description of the treatment, outcome and other independent variables

Variables	Description
private	if having a private health cover=1
visits.hopsital	if at least one visit to hospital=1
bmi	body mass index
income	individual income
age	individual age
education	years of education
health	5 levels: excellent=5, very good=6, good=7, fair=8, poor=9
race	4 levels: white=2, black=3, native American=4, other=5
limitation	if health puts binds on physical activity=1
region	4 levels: northeast=2, mid-west=3, south=4, west=5
gender	if male=1
hypertension	if hypertensive=1
hyperlipidemia	if hyperlipidemic=1
diabetes	if diabetic=1

measure the effect of the treatment (private) on the outcome (visit.hospital) variables; and hence, accounting for endogeneity, four types of copulas are proposed to be applied. The chosen copulas are: Gaussian, Frank, FGM and Plackett as in Section 3. For each copula, a new model is generated and the choice between these models will be made based on the Akike Information Criterion (AIC). Among all models, the AIC criterion selects the one which best fit the data. Table 7, summarizes the AIC and ATE for each copula. It is clear that the bivariate logit models, with all types of

Table 7: The AIC and ATE for the four types of compula for both logit and probit models

	logit		probit	
Type of Copula	$\widehat{ATE}$(CI)%	AIC	$\widehat{ATE}$(CI)%	AIC
Gaussian	1.59 (-1.90,5.43)	31583.66	1.02 (-5.72,5.65)	31763.88
Frank	3.29 (-2.98,9.02)	31584.71	3.51 (-4.69,11.31)	31764.81
FGM	0.419 (-9.263,8.379)	31584.39	3.66 (-1.91,9.70)	31764.82
Plackett	3.51 (-1.75,8.54)	31584.73	-0.38 (-7.93,7.92)	31764.44

copulas, have smaller AIC than bivariate probit models. Moreover, the Gaussian copula gives the minimum AIC for both logit and probit models; hence, the bivariate logit model with Gaussian copula is considered the best to fit this data; thus, estimating the dependence between the response variable (vist.hospital) and the treatment (private) more accuretly than othe models. The estimated ATE For the bivariate logit model with Gaussian copula, is equal to 1.59% which suggests that people who visits the hospital frequently are more likely to have a private health insurance by 1.59%. The simulated average effect is plotted in Figure 1.

The estimates of the model parameters with the Gaussian copula are listed in Table 8 below.

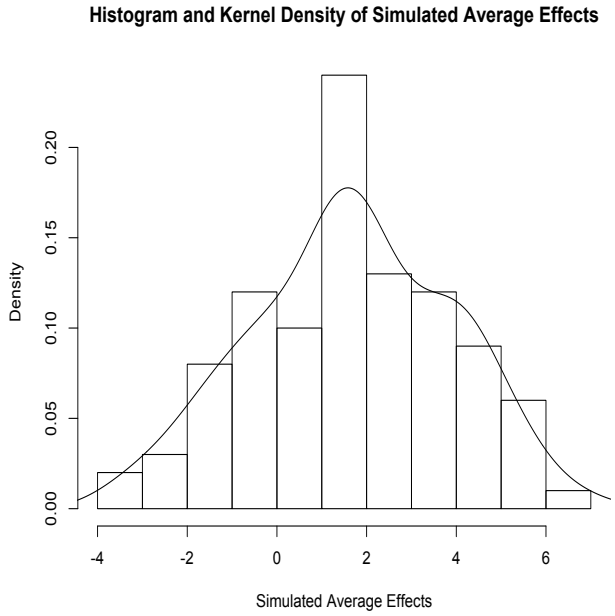

Figure 1: Plot of the histogram of simulated average effect together with the kernel estimate

Table 8: The estimates of the model parameters with the Gaussian copula

	Treatment equation		Outcome equation	
Variable	Estimate	Std.error	Estimate	Std.error
bmi	-0.001	0.003	0.004	0.003
private	—	—	0.154	0.233
income	2.2×10^{-5}	5.9×10^{-7}	8.6×10^{-7}	6.6×10^{-7}
age	0.020	0.001	0.022	0.002
education	0.228	0.007	0.064	0.012
health=6	-0.125	0.048	0.256	0.066
health=7	-0.269	0.051	0.326	0.069
health=8	-0.596	0.070	0.572	0.092
health=9	-0.966	0.111	0.945	0.130
race=3	-0.105	0.046	-0.202	0.062
race=4	-0.450	0.175	-0.171	0.240
race=5	0.132	0.071	-0.279	0.091
limitation	0.376	0.071	-0.787	0.074
region=3	0.459	0.062	0.269	0.070
region=4	0.122	0.054	-0.435	0.065
region=5	0.066	0.058	-0.693	0.0.074
gender	0.020	0.036	-0.707	0.047
hypertension	0.109	0.050	0.170	0.057
hyperlipidemia	0.256	0.051	0.450	0.054
diabetes	-0.019	0.073	0.189	0.075

5 Conclusions

Modeling the effect of a binary treatment on a binary response variable is of great importance. This research proposed a bivariate logit model to control endogeneity by estimating the dependence between a binary outcome and treatment variables by using copula function. A simulation study was conducted to measure the performance of different types of copula by measuring the ATE and MAPE under different sample sizes and four types of copula functions. The proposed

model was compared to the bivariate probit model via real applications under different copula functions, which show the potentiality of the proposed model.

Acknowledgments

The authors express their appreciation to the referees for thoroughly reviewing the details and providing valuable comments that enhanced the overall quality of the paper.

References

[1] L. Gitto, D. Santoro & G. Sobbrio, Choice of dialysis treatment and type of medical unit (private vs public): application of a recursive bivariate probit, *Health Economics*, **15**, 1251-1256 (2006)
[2] E. Latif, The impact of diabetes on employment in Canada, *Health Economics*, **18**, 577-589 (2009)
[3] G. Chen and S. Hamori, Bivariate probit analysis of differences between male and female formal employment in urban China, *Journal of Asian Economics*, **21**, 494-501 (2010)
[4] D. Goldman, J. Bhattacharya, D. McCaffrey, N. Duan, A. Leibowitz, G. Joyce and S. Morton, Effect of insurance on mortality in an HIV-positive population in care, *Journal of The American Statistical Association*, **96**, 883-894 (2001)
[5] H. Shelton Brown III, J. Pagán, and E. Bastida, The impact of diabetes on employment: genetic IVs in a bivariate probit, *Health Economics*, **14**, 537-544 (2005)
[6] A. Jones, X. Koolman, X and E. Van Doorslaer, The impact of having supplementary private health insurance on the use of specialists, *Annales D'Economie Et De Statistique*, **83-84**, 251-275 (2006)
[7] C. Fleming and P. Kler, I'm too clever for this job: a bivariate probit analysis on overeducation and job satisfaction in Australia, *Applied Economics*, **40**, 1123-1138 (2008)
[8] A. Kawatkar and M. Nichol, Estimation of causal effects of physical activity on obesity by a recursive bivariate probit model,*Value In Health*, **12**, A131-A132 (2009)
[9] Y. Li and G. Jensen, The impact of private long-term care insurance on the use of long-term care, *INQUIRY: The Journal Of Health Care Organization, Provision, And Financing*, **48**, 34-50 (2011)
[10] N. Daisy, M. Hafezi, L. Liu and H. Millward, Housing location and commuting mode choices of university students and employees: An application of bivariate Probit models, *International Conference on Transportation And Development*, 168-179 (2018)
[11] G. Marra, and R. Radice, A flexible instrumental variable approach, *Statistical Modelling*, **11**, 581-603 (2011)
[12] R. Radice, L. Zanin, and G. Marra, On the effect of obesity on employment in the presence of observed and unobserved confounding *Statistica Neerlandica*, **67**, 436-455 (2013)
[13] L. Zanin, R. Radice, and G. Marra, Modelling the impact of women's education on fertility in Malawi, *Journal Of Population Economics*, **28**, 89-111 (2015)
[14] F. Ieva, G. Marra, A. Paganoni and R. Radice, A semiparametric bivariate probit model for joint modeling of outcomes in STEMI patients, *Computational And Mathematical Methods In Medicine*, 1-7 (2014)
[15] G. Marra, G. Papageorgiou, and R. Radice, Estimation of a semiparametric recursive bivariate probit model with nonparametric mixing, *Australian & New Zealand Journal Of Statistics*, **55**, 321-342 (2013)
[16] R. Winkelmann, Copula bivariate probit models: with an application to medical expenditures, *Health Economics*, **21**, 1444-1455 (2012)
[17] R. Radice, G. Marra and M. Wojtyś, Copula regression spline models for binary outcomes, *Statistics And Computing*, **26**, 981-995 (2016)
[18] S. Han and E. Vytlacil, Identification in a generalization of bivariate probit models with dummy endogenous regressors, *Journal of Econometrics*, **199**, 63-73 (2017)
[19] T. Hasebe, On the treatment effects of a binary choice outcome model, *Economics Letters*, **200**, 109768 (2021)
[20] T. Hasebe, Endogenous models of binary choice outcomes: Copula-based maximum-likelihood estimation and treatment effects, *The Stata Journal*, **22**, 734-771 (2022)
[21] A. Abera and T. Abdisa, Financial inclusion and women's economic empowerment: Evidence from Ethiopia, *Cogent Economics & Finance*, **11**, 2244864 (2023)
[22] D. Shane, P. Trivedi and others, What drives differences in health care demand? the role of health insurance and selection bias, *Health, Econometrics and Data Group (HEDG) Working Papers*, **12**, (2012)