CHAPTER 2

Methodology

This chapter proposes econometrics models for research problems in health behaviors. The literature reviews on the copula theory and the switching regression model are discussed in Section 2.1 and Section 2.2 respectively.

2.1 Copula Modeling

Econometric estimation and inference for data that are assumed to be multivariate normal distributed are highly developed, but general approaches for joint nonlinear modeling of nonnormal data are not well developed, and there is a frequent tendency to consider modeling issues on a case-by-case basis.

Interest in copulas arises from several perspectives. First, econometricians often possess more information about marginal distributions of related variables than their joint distribution. The copula approach is a useful method for deriving joint distributions given the marginal distributions, especially when the variables are nonnormal. Second, in a bivariate context, copulas can be used to define nonparametric measures of dependence for pairs of random variables. When fairly general and/or asymmetric modes of dependence are relevant, such as those that go beyond correlation or linear association, then copulas play a special role in developing additional concepts and measures. Finally, copulas are useful extensions and generalizations of approaches for modeling joint distributions and dependence that have appeared in the literature.

One of the advantages of copula models is their relative mathematical simplicity. Another advantage is the possibility to build a variety of dependence structures based on existing parametric or non-parametric models of the marginal distributions.

2.1.1 Copula Function

We begin with the definition of copula, following Schweizer (1991). The theorem underlying of copula was introduced in 1959 by Abe Sklar. The theorem succinctly stated that an m-dimension copula (or m -copula) is a function C from the unit m-cube $[0,1]^{m}$ to the unit interval $[0,1]$ which satisfies the following conditions (Trivedi P.K. and Zimmer D.M., 2005):

(1) $C(1,\dots,1,a_n,1,\dots,1) = a_n$ for every $n \leq m$ and all a_n in $[0,1]$;

(2) $C(a_1, \dots, a_m) = 0$ if $a_n = 0$ for any $n \leq m$;

(3) C is m-increasing.

Property 1 says that if the realizations of $m-1$ variables are known each with marginal probability one, then the joint probability of the m outcomes is the same as the probability of the remaining uncertain outcome. Property 2 is that the joint probability of all outcomes is zero if the marginal probability of any outcome is zero. Property 3 says that the C-volume of any m-dimensional interval is non-negative. Properties 2 and 3 are general properties of multivariate cdfs that were previously mentioned.

It follows that an m-copula can be defined as an m-dimensional cdf whose support is contained in $[0,1]^{m}$ and whose one-dimensional margins are uniform on [0,1]. In other words, an m-copula is an m-dimensional distribution function with all m univariate margins being U(0,1). To see the relationship between distribution functions and copulas, consider a continuous m-variate distribution function $F(y_1,\dots,y_m)$ with univariate marginal distributions $F_1(y_1),\dots,F_m(y_m)$ and inverse functions $F_1^{-1},\dots,F_m^{-1}$. Then $y_1=F_1^{-1}(u_1) \sim F_m$ where $u_1,\dots u_m$ are uniformly distributed variates. The transforms of uniform variates are distributed as F_i (i = 1,…,m). Hence

$$\begin{aligned} F(y_1,\dots,y_m) &= F\left(F_1^{-1}(u_1),\dots,F_m^{-1}(u_m)\right) \\ &= \Pr[U_1 \leq u_1,\dots,U_m \leq u_m] \\ &= C(u_1,\dots,u_m) \qquad (2.1) \end{aligned}$$

is the unique copula associated with the distribution function.

For an m-variate function F, the copula associated with F is a distribution function C: $[0,1]^{m}$ →[0,1] that satisfies

$$F(y_1, \dots, y_m) = C(F_1(y_1), \dots, F_m(y_m); \theta) \tag{2.2}$$

where θ is a parameter of the copula called the dependence parameter, which measures dependence between the marginals. $\mathbf{y} = (y_1,..., y_m)$ is the realization of an m-dimensional random vector $\mathbf{Y} = (Y_1,...,Y_m)$. $F_j(y_j)$ is the marginal distribution function of the j^{th} margin for $j = 1, \dots, m$ and F is a joint distribution function.

For continuous variables, the joint density $f(y_1,..., y_m)$ can be easily obtained by taking the derivative of both the sides of equation (2.2), which gives

$$f(y_1,..., y_m) = c(F_1(y_1),..., F_m(y_m)) f_1(y_1)... f_m(y_m),$$

where f is a joint density function, f_j is a marginal density function corresponding to each marginal j , and c is a copula density function. The copula function is unique for the continuous random vector $\mathbf{Y}$. However, the copula function is unique only over the Cartesian product of the ranges of the marginal distribution function in a discrete random vector (Genest and Neslehova, 2007). For modeling issues, parametric modeling of discrete variables by copula acquires dependence properties in the same way as in the continuous case.

For discrete variables, the probability mass function can be evaluated by taking the difference of the copula function. The joint probability mass function (pmf) of $\mathbf{Y}$ can be obtained as follows:

$$\Pr(\mathbf{Y} = y) = \sum_{i_1=0,1} \cdots \sum_{i_m=0,1} (-1)^{i_1+\dots+i_m} C(F_1(y_1 - i_1), \dots, F_m(y_m - i_m)) \tag{2.3}$$

Note that, to compute this pmf, we have to evaluate 2^m times of the copula functions. However, one can approximate the pmf of $\mathbf{Y}$ by building up from the number of bivariate copulas. This approach is called pair copula constructions (PCC).

2.1.2 Pair Copula Constructions

Pair copula constructions (PCC) were initiated by Joe (1996) and developed in more detail by Bedford and Cook (2001, 2002), and Kurowicka and Cooke (2006).

For continuous $\mathbf{Y}$, a PCC can be derived by factorizing the joint density function into the conditional density function and the marginal density function, as follows:

$$f(y_1,...,y_m) = f_{1|2,...,m}(y_1|y_2,...,y_m)\, f_{2|3,...,m}(y_2|y_3,...,y_m)...f_m(y_m) \tag{2.4}$$

Aas et al. (2009) have shown that the conditional density function on the right hand side of equation (2.4) can be decomposed into the product of a bivariate copula density and a univariate conditional density by using Sklar's theorem. This can be done recursively to each of the terms on the right hand side of equation (2.4) until $f(y_1,...,y_m)$ is decomposed into the product of m(m-1)/2 bivariate copulas (Panagiotelis, 2012).

For discrete margins, we can decompose a pmf by using the method proposed by Panagiotelis, 2012 as follows:

$$\begin{aligned}\Pr(Y_1 = y_1,...,Y_m = y_m) = \Pr(Y_1 = y_1|Y_2 = y_2,...,Y_m = y_m)\times \\ \Pr(Y_2 = y_2|Y_3 = y_3,...,Y_m = y_m)\times...\times\Pr(Y_m = y_m)\end{aligned} \tag{2.5}$$

We can perform the same decomposition as in a continuous case for each term on the right hand side of equation (2.5) to get the product of a bivariate copula.

For example in the case of m = 3, three-dimensional discrete margin PCC can be obtained as follows:

$$\begin{aligned}&\Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \\ &\quad = \Pr(Y_1 = y_1|Y_2 = y_2, Y_3 = y_3)\times\Pr(Y_2 = y_2|Y_3 = y_3)\times\Pr(Y_3 = y_3),\end{aligned} \tag{2.6}$$

where

$$\Pr(Y_1 = y_1 | Y_2 = y_2, Y_3 = y_3) = \frac{\left\{ \sum_{i_1=0,1} \sum_{i_2=0,1} (-1)^{i_1+i_2} C_{12|3}(F_{1|3}(y_1 - i_1 | y_3), F_{2|3}(y_2 - i_2 | y_3)) \right\}}{\Pr(Y_2 = y_2 | Y_3 = y_3)}, \tag{2.7}$$

and the arguments in the copula function are

$$F_{1|3}(y_1 - i_1 | y_3) = \frac{C_{13}(F_1(y_1 - i_1), F_3(y_3)) - C_{13}(F_1(y_1 - i_1), F_3(y_3 - 1))}{\Pr(Y_3 = y_3)} \ ,$$

and

$$F_{2|3}(y_2 - i_2 | y_3) = \frac{C_{23}(F_2(y_2 - i_2), F_3(y_3)) - C_{23}(F_2(y_2 - i_2), F_3(y_3 - 1))}{\Pr(Y_3 = y_3)} .$$

Since the dominator of equation (2.7) cancels with the second term on the right hand side of equation (2.6), the full expression for the pmf of the three-dimensional discrete margin PCC is

$$\begin{aligned}\Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \\ = \left\{ \sum_{i_1=0,1} \sum_{i_2=0,1} (-1)^{i_1+i_2} C_{12|3} \left(\frac{C_{13}(F_1(y_1 - i_1), F_3(y_3)) - C_{13}(F_1(y_1 - i_1), F_3(y_3 - 1))}{F_3(y_3) - F_3(y_3 - 1)}, \right. \right. \\ \left. \left. \frac{C_{23}(F_2(y_2 - i_2), F_3(y_3)) - C_{23}(F_2(y_2 - i_2), F_3(y_3 - 1))}{F_3(y_3) - F_3(y_3 - 1)} \right) \right\} \left[F_3(y_3) - F_3(y_3 - 1) \right]\end{aligned}$$

The above model can be used to analyzed the dependence between each health behavior considered in this paper and can also determine the factors affecting those behaviors as the same time.

2.1.3 Estimation and Model Selection

This thesis, we estimate both the copula and marginal parameters jointly by the maximum likelihood estimator (MLE) of the model parameters involving simultaneous maximization of the log-likelihood over the dependence (θ) and marginal parameters.

We determine the best model after fitting different PCC models to a given data set, one can rely on the classical AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

2.2 Switching Regression Model for Level of Hypertension

In the sample used to estimate the effect of alcohol consumption on the level of hypertension, the participants are not randomly drawn from the population from which we wanted to draw inferences, but participants who self-selected themselves into treatment. The approach to self-selection used, is that proposed by Heckman (Heckman, 1979). The assumption is that the self-select mechanism may be modeled by a binary choice model. The switching regression model, was supplemented with copula by using copula to model the correlation between the random errors from a decision model and outcome models (Heckman, 1979).

Consider two decisions, S=0,1 , where 1 is 'drink alcohol' and 0 is not. Let $S^* = Z\gamma + \nu$ be the latent variable for the decision mechanism. The decision rule is the following condition

$$S = \begin{cases} 1 \;\; if\; s^* > 0 \\ 0 \;\; if\; s^* \leq 0 \end{cases}$$

where Z is the matrix of the explanatory variables explaining the self-select mechanism, and γ is the corresponding vector of parameters to be estimated. The individuals are observed either in decision S=0 , or in decision S=1, but never in both.

Consider the outcome of interest, the level of hypertension *Ys = 0,...,J,* can be modeled using the latent variable framework and can be determined by the following condition:

$$Y_s = j \;\; iff \;\; k_{s,j-1} < Y_s^* \leq k_{s,j}, \qquad S = 0,1, \qquad j = 0, \dots, j \qquad (1)$$

where $\kappa_{s,j}$ are the threshold values, which form a partition of the real line, i.e.,$\kappa_{s,0} = -\infty$, $\kappa_{s,J} = \infty$, and $\kappa_{s,j} > \kappa_{s,j-1}$ for all j.

Let $Y_0^* = X\beta_0 + \varepsilon_0$ be the latent variable for the individual decision not to drink S = 0 , and $Y_1^* = X\beta_1 + \varepsilon_1$ be the latent variable for the individual to consume alcohol S = 1, where, X is the vector of all the explanatory variables, β_0 and β_1 are the vector of the parameters to be estimated.

As previously discussed, there might be some unobserved factors affecting both the self-selected mechanism and the response outcome, therefore the probability of observing *Ys* = *j* depends on the self-selected variable S. Given that *S* and *Ys* are not necessarily independent. We have

$$\Pr(Y_0 = j, S = 0 \mid X, Z) = P(\kappa_{0,j-1} - X\beta_0 < \varepsilon_0 \le \kappa_{0,j} - X\beta_0, \nu \le -Z\gamma)$$

$$= P(\varepsilon_0 < \kappa_{0,j} - X\beta_0, \nu \le -Z\gamma) - P(\varepsilon_0 < \kappa_{0,j-1} - X\beta_0, \nu \le -Z\gamma)$$

$$\Pr(Y_1 = j, S = 1 \mid X, Z) = P(\kappa_{1,j-1} - X\beta_1 < \varepsilon_1 \le \kappa_{1,j} - X\beta_1, \nu > -Z\gamma)$$

$$= P(\varepsilon_1 < \kappa_{1,j} - X\beta_1) - P(\varepsilon_1 < \kappa_{1,j-1} - X\beta_1)$$

$$- P(\varepsilon_1 < \kappa_{1,j} - X\beta_1, \nu \le -Z\gamma) + P(\varepsilon_1 < \kappa_{1,j-1} - X\beta_1, \nu \le -Z\gamma)$$

To model the above probability, we have to specify the appropriate joint distribution functions. In this chapter, we suggest combining the marginal distributions (ε_s and ν) by using copula.

For a bivariate joint distribution H with marginal distributions F_1 and F_2, the copula C : $[0,1]^2 \rightarrow [0,1]$, which combines these two marginal distributions, can be expressed as follows:

$$H(x, y) = C\{F_1(x_1), F_2(x_2)\},\ (x, y) \in R^2$$

The copula function is uniquely determined for the continuous random vector (F_1,F_2). For a discrete random vector, the copula function is unique only over the Cartesian product of the range of the marginal distribution function (Genest and Neslehova 2007). Thus, in discrete cases the mapping from two marginal distributions and copula to a bivariate joint distribution is not one-to-one. However, the region outside the Cartesian product of the range of the marginal distribution function is not of interest (Nelsen 2006). Moreover, Genest, C. and Neslehova, J. demonstrated that parametric modeling of discrete random vector by copula acquires dependence properties in a way that is similar to the continuous case (Genest and Neslehova 2007).

For any copula, the marginal distribution implied by bivariate copula are C(u,v) ≤ C(u,1) = u and C(u,v) ≤ C(1,v) = v, for all 0 ≤ u,v < 1 , and so W(u,v) = max(u+v-1, 0) ≤ C(u,v) ≤ min(u,v)=M(u,v). The copula M(u,v) and W(u,v) are called the Frechet upper bound and Frechet lower bound, respectively. We can interpret the Frechet lower bound as the copula with the maximum negative dependence and Frechet upper bound as the copula with the maximum positive dependence. In modeling switching regression, it is essential that the copula should allow for both positive and negative dependence, since the direction of the selection bias can be in both directions. We should not restrict the direction of selection bias a priori. The selection pattern should be explained by the data itself.

Copula has had limited use in the endogenous switching regression models. Some, but not all examples, are (Sirisrisakulchai & Sriboonchitta, 2014a) and (Ophem, 2000) for modeling endogenous switching regression in count outcomes, (Smith, 2005) for modeling endogenous switching regression of continuous variables and (Luechinger et. al, 2010) for modeling endogenous switching regression in ordered outcomes.

For any given copula, the two required joint distribution, Pr(Y_0 = j, S = 0 |X, Z) and Pr(Y_1 = j, S=1 |X, Z) are fully determined. Therefore,

$$\Pr(Y_0 = j, S = 0 | X, Z) = C_0(F_1(\kappa_{0,j} - X\beta_0), F_2(-Z\gamma); \theta_0) - C_0(F_1(\kappa_{0,j-1} - X\beta_0), F_2(-Z\gamma); \theta_0)$$

and

$$\Pr(Y_1 = j, S = 1 | X, Z) = C_1(F_1(\kappa_{1,j} - X\beta_1), 1; \theta_1) - C_1(F_1(\kappa_{1,j-1} - X\beta_1), 1; \theta_1)$$

$$- C_1(F_1(\kappa_{1,j} - X\beta_1), F_2(-Z\gamma); \theta_1) + C_1(F_1(\kappa_{1,j-1} - X\beta_1), F_2(-Z\gamma); \theta_1)$$

where C_0(u,v) and C_1(u,v) are copula functions and F_1 and F_2 are marginal functions which can be either normal or logistic distribution which correspond to the Probit and Logit models, respectively.