Thai Journal of Mathematics (2014) 91–110
Special Issue on : Copula Mathematics and Econometrics
http://thaijmath.in.cmu.ac.th
Online ISSN 1686-0209

Econometric Analysis of Older of Older Workers' Hours of Work Using A Copula and Sample Selection Approach

A.Wichian[†1] **and S.Sriboonchitta**[‡],

[†]Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand
e-mail : aunwichian@hotmail.com
[‡]Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand
e-mail : songsakecon@gmail.com

Abstract : This paper aims at a copula approach to a sample selection model to determine the factors affecting the labor force participation and working hours of older workers in Thailand. Several of the copula functions, such as the Gaussian, and Archimedean copulas are compared to the standard Heckman's method which is restricted to linear dependence. This paper demonstrates that the copula approach to the sample selection model, which allows for flexibility in dependence structure and relaxing the joint normality assumption, works in this context and performs better than the BVN (bivariate normal model) especially, the Frank (L–t) copula model. Furthermore, the results obtained show the presence of significant positive dependency of unobserved factors between two error terms, which implies that the selectivity bias exists and also that the unobserved factors that increase (decrease) the propensity to participate in the labor force of older workers also increase (decrease) the working hours. In addition, the policy recommendation is that the stakeholders should consider on having a program that can increase working hours of some worker groups such as female workers, or workers who are not single or workers who do not stay in the Bangkok province.

Keywords : sample selection; copula; older workers
2010 Mathematics Subject Classification : 62P20; 91B84 (2010 MSC)

---

[1]Corresponding author.

1 Introduction

Labor supply is a key component in economic growth, especially, in the labor force participation of workers aged 15–59. However, at the present, the demographic structure is undergoing a change because of an aging society. In Thailand, between 2000 and 2010, the proportion of people aged 0–14 shows that it is declining, while the proportion of people aged 60 and above shows an increase. The proportion of the latter was 12.0 percent in 2011 (National Statistical Office [1]). These data point out that in the future, older workers will inevitably become significant contributors to Thailand's economic activity. Nevertheless, the Labor Force Survey of Whole Kingdom found that the older worker participation rate was only 38.6 percent (National Statistical Office [1]). Therefore, an increase in the participation rate of older workers should be taken into consideration, especially when the country is on the verge of becoming an absolutely aging society. Also, in order to increase the participation rate, the predominant questions about the factors that determine the participation decisions and working hours should be identified. Thus, the purpose of this paper is to investigate the determinants of older worker participation decisions and working hours in Thailand. The results will be useful for policy makers to adjust the policies that can stimulate the participation of older workers.

Since not all older workers participate in the labor market, it is inevitable that the study of participation decisions is associated with sample selection problems.

The sample selection model has been widely used in the microeconomics field for a long time. Since economic applications have frequently encountered the nonrandom sampling, this causes selectivity bias, which leads to the inconsistency of the OLS estimator (see Wooldridge [2, p. 560–563]). Thus, the econometricians have tried to correct the sample selection bias by using various methods. Two correction methods have been proposed by Heckman [3], [4]: the first is full information maximum likelihood method (FIML), in 1974, and the second is Heckman's [4] two-step estimation. The latter is usually used instead of the former (see Greene [5, p.784]; Vella [6]). Nevertheless, the latter method of estimation may address the collinearity problem between the regressors in the selection equation and the outcome equation (see Puhani [7]). In addition, most empirical studies point out that the ML estimator is more efficient and outperforms when it is compared to the two-step estimator (see Nawata [8]; Nawata and Li [9]). However, the ML estimator has some crucial drawbacks, namely, it has strong assumptions of bivariate normality for the joint distribution, which leads to incorrect conclusion about the existence of sample selection bias. Therefore, econometricians have tried to find the best procedure which can relax the above assumptions and attain the robust estimators. A particular mainstream procedure used nonparametric or semi-parametric methods to relax this strong assumption. Unfortunately, this method has some significant drawbacks. For example, it cannot identify the intercept in the outcome equation (see Genius and Strazzera [10]). In addition, the difficulty in implementation and in the estimation of the associated covariance matrices required for inference causes the semi-parametric methods to have been less frequently employed in empirical work (see Vella [6]). Another procedure that still used FIML, but relaxes the normality assumption of the marginal, is the model had been proposed by Lee [11], [12]. Lee allowed the marginal distribution of the disturbances to be non–normal, and then transformed it into normal distribution. But he still used the bivariate normal distribution to capture the dependence between the transformed disturbances. Although Lee tried to avoid the strong assumption, this method still maintains the bivariate normal distribution, which implies linear dependence between the disturbances.

In recent times, the copula approach has been widely used in the sample selection framework. Trivedi and Zimmer [13] pointed out the many reasons that lead to an interest in the copulas. First, the joint distributions can be derived when the marginal distributions are given, especially for non–normal margins. Second, the concepts and measures of dependence that goes beyond correlation or linear association can be developed. The copula approach is very useful for the topics that relate to the joint distribution as the sample selection framework. Thus, when the researchers apply the copula approach to the sample selection framework, not only can they allow different specifications for the marginal, but also use wide ranges of distribution shapes, both symmetric and asymmetric. Actually, Lee's [12] method is among the early work in the copula framework, but the term "copulas" is not explicitly used (see Trivedi and Zimmer [13]). Nonetheless, the properties of the copulas have obviously been employed in Smith's [14] paper which suggests the general form for the self–selection model using the properties of copula to measure the dependence of the disturbances in the FIML. There are several papers which have applied the copula approach and followed Smith's [14]procedure (for example, Genius and Strazzera [10]; Eberth and Smith [15]; Hasebe and Vijverberg [16]; Chinnakum, Sriboonchitta and Pastpipatkul [17];Sirisrisakulchai and Sriboonchitta [18] etc.). Although these studies have the same framework for copula sample selection, the marginal distribution and the joint distribution that have been used are different, depending on the empirical distribution of the data. Since the researcher does not have prior knowledge about the dependence structure, most studies have tried to test the various copulas functions, except with the work by Dancer et al. [19] which used only Gaussian and Frank copulas. Eberth and Smith's study [15], which used only the Gaussian copula. Consider the marginal distributions of the error terms. Dancer et al. [19], Bhat and Eluru [20], and Chinnakum, Sriboonchitta and Pastpipatkul [17] used normal distribution. However, the marginal distributions were not needed simultaneously. For examples, Genius and Strazzera [10] estimated models based on different marginal distributions such as the logistic and the Student–t_v distributions. Sener and Bhat [21] used four possible combinations for the marginal distribution, namely, normal–normal, logistic–logistic, normal–logistic and logistic–normal. On the other hand, Hasebe and Vijverberg [16] argued that it is not usually known as a priori on marginal distribution. Thus they proposed a new flexible distribution, Generalized Tukey lambda (GTL) which is used for each margin.

Finally, the copula approach is very useful for the context where one does not have prior knowledge about the dependence structure that holds between two unobserved factors, and to relax the strong assumptions of bivariate normality for the joint distribution. Therefore, in this current paper, we try to show that the copula approach to sample selection works and performs better than the traditional method in the context of labor participation. To attain this purpose, we apply the copula approach to a sample selection modeling to determine the factors affecting the labor force participation and working hours of older workers in Thailand. Importantly, we attempt to relax the joint normality assumption and test different functional form of copula that are both radially symmetric and asymmetric copulas, such as, the Gaussian, FGM, and Archimedean copulas (AMH, Clayton, Frank, Gumbel and Joe copulas) in order to compare with the results of the standard Heckman's method (or bivariate normal model, BVN), which is restricted to bivariate normal distributional assumption. To our knowledge, this is the first application of a copula framework to focus on a context of labor participation in Thailand.

This paper is organized as the following: Section 2 describes the copula theory, the definitions, and the main properties, bounds of copulas, related measures of dependence, and some examples of copula. Section 3 describes the sample selection model as well as on explaining how to apply the copula functions to the sample selection model. Section 4 describes the data. Section 5 is devoted to the application of the copula sample selection model to investigate the factors that determine the working hours of older workers. Finally, section 6 provides the conclusion for the work.

2 Copula Theory

2.1 Definition and Properties

The term "copula" is defined as functions that link or connect multivariate distributions to their one-dimensional margins (see Schweizer [22]; Trivedi and Zimmer [13]). We begin with a bivariate copula function, which is defined as follows (see Nelsen [23, p.10]):

Definition: A copula is a function $C : [0,1]^2 \to [0,1]$ with the following properties

1. For every u, v in $[0,1]$,
$C(u,0) = 0 = C(0,v)$ and $C(u,1) = u$ and $C(1,v) = v$

2. For every u_1, u_2, v_1, v_2 in [0,1] such that $u_1 \leqslant u_2$ and $v_1 \leqslant v_2$,
$C(u_2,v_2)–C(u_2,v_1)–C(u_1,v_2)+C(u_1,v_1) \geqslant 0$.

Essentially, the theoretical foundation is provided by Sklar's theorem, as given below (see Nelsen [23, p. 18]):

Sklar's theorem: Let X and Y be random variables and H be a joint distribution function with margins F and G, which are the cumulative distribution functions of the random variables X and Y, respectively. Then, there exists a copula C such that for all x, y in $\overline{R}$.

$$H(x,y) = C(F(x), G(y)) \tag{2.1}$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $RanF \times RanG$. Conversely, if C is a copula and F and G are distribution functions, then the function H defined by (2.1) is a joint distribution function with margins F and G.

By Sklar's theorem and the method of inversion, the corresponding copula can be generated by using the unique inverse transformations $x = F^{-1}(u)$ and $y = G^{-1}(v)$. Therefore,

$$C(u,v) = H(F^{-1}(u), G^{-1}(v)), \tag{2.2}$$

where u and v are standard uniform variates.

In practical implications, copulas allow researchers to piece together joint distributions when only marginal distributions are known with certainty. For a two–variate function with margins F and G, the copula associated with H is a distribution function $C : [0,1]^2 \to [0,1]$ that satisfies

$$H(x,y) = C(F(x), G(y); \theta), \tag{2.3}$$

where θ is a parameter of the copula called the dependence parameter, which measures the dependence between the marginals (see Trivedi and Zimmer [13]).

Furthermore, the dependence parameter can be used to denote the families of the copulas as notation $C_\theta(u,v)$. There are several examples of families of copulas, such as the Gaussian (Normal) copula, the FGM (Farlie–Gumbel–Morgenstern) copula, the Plackett copula, etc.

In recent times, the bivariate copula has been widely used in several of economics research fields such as financial economics (for examples: Patton [24]; Boonyanuphong and Sriboonchitta [25] etc.), tourism economics (for examples Puarattanaarunkorn and Sriboonchitta [26] etc.) and agricultural economics (for examples Sriboonchitta et al [27]; Xue and Sriboonchitta [28] etc.).

2.2 Bounds of Copula

The additional property that relates to copulas is the Fréchet–Hoeffding bounds. Application of the Fréchet–Hoeffding bounds to a copula in the bivariate case, for any copula C and for all u, v in [0,1], is given by

$$W(u,v) = max(u+v-1, 0) \leqslant C(u,v) \leqslant min(u,v) = M(u,v), \tag{2.4}$$

where W is the Fréchet–Hoeffding lower bound, and M is the Fréchet–Hoeffding upper bound. Usually, the copula lies between these bounds. The copula attaining the the Fréchet–Hoeffding lower bound corresponds to negative dependence. The copula attaining the Fréchet–Hoeffding upper bound corresponds to positive dependence. And, in special cases of copulas, the product copula can be defined if the margins are independent (see Schmidt [29]; Trivedi and Zimmer [13]). In addition, the families of copulas are called comprehensive if they include both Fréchet–Hoeffding bounds and product copula.

The copulas such as the Gaussian and the Frank copulas are comprehensive, while the FGM, Clayton, Gumbel, and Joe copulas are not comprehensive, which makes it necessary to calculate the measures of dependence, as described below.

2.3 Measures of Dependence

As mentioned above, some families of copulas are not comprehensive; thus, in such cases the measure of dependence can be used to assess the coverage of the copula. The measures of dependence can be determined using several alternative methods. The most familiar and often used method is the linear correlation, such as the Pearson's product moment correlation coefficient. But this measure has some drawbacks: first, in general zero correlation, it does not imply independence. Second, it is not defined for heavy–tailed distributions whose second moments do not exist. Third, it is not invariant under strictly increasing non–linear transformations (see Trivedi and Zimmer [13]). Therefore, the alternative methods to compute measures of dependence have come up–such as the concordance measures–which the statistician usually uses as Kendall's τ and Spearman's ρ_S. The former is defined as follows:

$$\tau = P((X - X')(Y - Y') > 0) - P((X - X')(Y - Y') < 0), \tag{2.5}$$

and the latter is defined as follows:

$$\rho_S = 3(P((X - X')(Y - Y'') > 0) - P((X - X')(Y - Y'') < 0)), \tag{2.6}$$

where (X, Y),(X', Y'),and (X'', Y'') are independent random vectors, and each vector has a joint distribution function $F(..., ...)$ whose margins are F_1 and F_2.

Since (X,Y) are continuous random variables whose copula is $C_\theta(u, v)$, the Kendall's τ can be expressed in terms of copulas (see Nelson [30, p.129]):

$$\tau = 4\int\int_{[0,1]^2} C_\theta(u, v)dC_\theta(u, v) - 1 = 4E(C_\theta(U, V)) - 1, \tag{2.7}$$

where the second expression is the expected value of the function $C_\theta(U, V)$ of uniform (0,1) random variables U and V with a joint distribution function C.

Also, Spearman's ρ_S can be simplified thus, in terms of copulas:

$$\rho_S = 12\int\int_{[0,1]^2} uvdC_\theta(u, v) - 3 = 12E(UV) - 3, \tag{2.8}$$

where $U = F(X)$ and $V = F(Y)$ are uniform (0,1) random variables with joint distribution function $C_\theta(u, v)$. Both of the concordance measures are bounded between −1 and 1, and zero under the product copula.

2.4 Some Bivariate Copulas

There are several examples of copulas, such as product copulas, Farlie–Gumbel–Morgenstern (FGM) copulas, Gaussian copulas, and the Archimedean class of copulas, which are as defined:

2.4.1 Product copulas

Product copulas have the form

$$C(u, v) = uv, \tag{2.9}$$

where (u, v) is in $[0, 1]^2$. This copula is in correspondence with the independence of the random variables.

2.4.2 Gaussian copulas

The Gaussian, or Normal copula, was proposed by Lee [12] for selectivity models, and is given by

$$C_\theta(u, v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta), \tag{2.10}$$

where $\Phi(.)$ is the cumulative distribution function of a standard normal variate, and $\Phi_2(.)$ is the bivariate cumulative distribution function with Pearson's correlation parameter $\theta(-1 \leqslant \theta \leqslant 1)$. The Gaussian copula is comprehensive since it includes the product copula and both of the Fréchet–Hoeffding bounds, and captures both positive and negative dependences. However, it is radially symmetric in its dependence structure and strong central dependency. The concordance measures for the Gaussian copula can be given in terms of the dependence parameter (θ) as $\tau = (2/\pi)sin^{-1}(\theta)$ for Kendall's τ and $\rho_S = (6/\pi)sin^{-1}(\theta/2)$ for Spearman's ρ_S

2.4.3 The Farlie–Gumbel–Morgenstern (FGM) copulas

This copula has a simple form, which is given by

$$C_\theta(u, v) = uv(1 + \theta(1 - u)(1 - v)), \tag{2.11}$$

where $-1 \leqslant \theta \leqslant 1$.

If the dependence parameter(θ) equals zero, then it leads to the product copula. Although the FGM copula has a simple form, it is not comprehensive because it includes only the product copula, and not the Fréchet–Hoeffding lower and upper bounds. This copula is radially symmetric in its dependence structure to the Gaussian copula, but the dependence structure is weaker than that of the Gaussian. In addition, this copula is only useful in cases of moderate dependency (see Trivedi and Zimmer [13]).The concordance measures for the FGM copula can be given in terms of the dependence parameter (θ) as $\tau = \frac{2}{9}\theta$ for Kendall's τ and $\rho_S = \frac{1}{3}\theta$ for Spearman's ρ_S.

2.4.4 The Archimedean copulas

The Archimedean copulas are popular in empirical works for several reasons. These copulas can display a wide range of dependence properties for different choices of generator function (see Trivedi and Zimmer [13]). Furthermore, Smith [14] pointed out that it makes estimation of the maximum likelihood and calculation of the score function relatively easy. In order to better understand the Archimedean copulas, we need to mention some properties of these copulas. The bivariate Archimedean copulas can be generated in the following form:

$$C_\theta(u, v) = \varphi^{-1}[\varphi(u) + \varphi(v)], \tag{2.12}$$

where $\varphi : [0, 1] \to [0, \alpha]$ is a generator function which satisfies the following properties: $\varphi(1) = 0, \varphi'(t) < 0$, and $\varphi''(t) > 0$ for $0 < t < 1$. In addition, if $\varphi(0) = \alpha$, then the inverse function φ^{-1}exists.

The above form can be written as follows:

$$\varphi(C_\theta(u, v)) = \varphi(u) + \varphi(v), \tag{2.13}$$

Taking the differential with respect to v in the above equation, we obtain the result which will be used in the sample selection model, which can be given as

$$\frac{\partial(C(u, v))}{\partial(v)} = \frac{\varphi'(v)}{\varphi'(C(u, v))}, \tag{2.14}$$

Also, for the Archimedean copula, Kendall's τ can be described in simple form, as follows:

$$\tau = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt, \tag{2.15}$$

where $\varphi'(t) = \partial\varphi(t)/\partial(t)$; The Clayton, Frank, Gumbel and Joe copulas are the Archimedean copulas, and they have been extensively used in empirical work (for example, Genius and Strazzera [10]; Sener and Bhat [21]; Hasebe and Vijverberg [16]; Chinnakum, Sriboonchitta and Pastpipatkul [17]). These copulas are different in the generate function, which leads to a difference in the functional form (which is demonstrated in Table 1) and an essential dependence structure. A description is provided below.

Table 1: Functional Form and Characteristics of Bivariate Copulas

Copula	Function C(u,v)	Generation function	Range of θ	Range of Kendall's τ
Gaussian	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta)$	-	$-1 \leqslant \theta \leqslant 1$	$-1 \leqslant \tau \leqslant 1$
FGM	$uv(1+\theta(1-u)(1-v))$	-	$-1 \leqslant \theta \leqslant 1$	$-2/9 \leqslant \tau \leqslant 2/9$
AMH	$uv/(1-\theta(1-u)(1-v))$	$log\frac{1-\theta(1-t)}{t}$	$-1 \leqslant \theta \leqslant 1$	$-0.18 \leqslant \tau < 1/3$
Clayton	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$(1/\theta)(t^{-\theta}-1)$	$0 < \theta < \alpha$	$0 < \tau < 1$
Frank	$-\frac{1}{\theta}ln\{1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\}$	$-ln[(e^{\theta t}-1)(e^{\theta}-1)]$	$-\alpha < \theta < \alpha$	$-1 \leqslant \tau \leqslant 1$
Gumbel	$exp(-[(-lnu)^{\theta}+(-lnv)^{\theta}]^{1/\theta})$	$(-lnt)^{\theta}$	$1 \leqslant \theta < \alpha$	$0 \leqslant \tau < 1$
Joe	$1-[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}]^{1/\theta}$	$-ln[1-(1-t)^{\theta}]$	$1 \leqslant \theta < \alpha$	$0 \leqslant \tau < 1$

Source: The copula functions are given as presented in Trivedi and Zimmer [13] and Smith [14]

Frank copula

The first example of the Archimedean copulas is the Frank copula, which was proposed by Frank [31]. This copula attains the Fréchet–Hoeffding upper and lower bounds and corresponds to independence, and so it is comprehensive as a Gaussian copula. This property implies that this copula can account for both positive and negative dependences. Furthermore, this copula is radially symmetric to Gaussian and FGM copulas, but has stronger central dependence than the Gaussian copula. Thus, Frank copula is most appropriate for data that exhibit very strong central dependence and weak tail dependence. (see Trivedi and Zimmer [13]; Bhat and Eluru [20]). Kendall's τ may be taken in the form given below (see Nelson [13, p.171]):

$$\tau = 1 - \frac{4}{\theta}[1 - D_F(\theta)], D_F(\theta) = \frac{1}{\theta}\int_{t=0}^{\theta} \frac{t}{e^t - 1}dt, \tag{2.16}$$

The range of τ is $-1 \leqslant \tau \leqslant 1$.

Clayton copula

The second example is the Clayton copula, which was proposed by Clayton [32]. This copula is not comprehensive since it only attains the Fréchet–Hoeffding upper bound and corresponds to independence when $\theta \to 0$, which implies that it cannot account for negative dependence. In addition, this copula exhibits asymmetry in the sense that there is a strong left tail dependence. Kendall's τ can be given in the simple form of the dependence parameter (θ) as $\tau = \theta/(\theta + 2)$. Therefore, τ is restricted on the region (0,α) or $0 < \tau < 1$.

Joe copula

Another example of the Archimedean copulas is the Joe copula, which was proposed by Joe [33]. This copula exhibits asymmetry and cannot account for negative dependence, just like the Clayton copula. However, it is opposite to the Clayton copula in dependence, that is, there is a strong right tail dependence. The

Kendall's τ may be taken in the form that is given below:

$$\tau = 1 + \frac{4}{\theta} D_J(\theta), D_J(\theta) = \int_{t=0}^{1} \frac{[ln(1-t^\theta)](1-t^\theta)}{t^{\theta-1}} dt, \tag{2.17}$$

The range of τ is $0 \leqslant \tau < 1$, and corresponds to independence as $\theta = 1$.

The characteristics for each of the copulas are given in Table 1.

3 Copula Approach to Sample Selection Model

3.1 Sample Selection Model

The sample selection model has several forms; however, in this paper, we analyze the form that Amemiya [34] called the type 2 Tobit model (see Amemiya [34]), for the general framework) which takes the form that relates to this context, and is as given below:

$$\begin{aligned} y_i^* &= z_i^{'}\gamma + u_i, y_i = \mathbf{1}[y_i^* > 0] \text{ selection equation} \\ h_i^* &= x_i^{'}\beta + \varepsilon_i, h_i = \mathbf{1}[y_i^* > 0]h_i^* \text{ outcome equation} \end{aligned} \tag{3.1}$$

The notation $\mathbf{1}[y_i^* > 0]$ denotes the indicator function, which takes the value 1 if event $y_i^* > 0$ and the value 0 if otherwise. The selection equation represents the binary decision of older workers regarding the question of whether to participate in the labor force or not, and it is equal to 1 if the older worker decides to participate in the labor force and equal to 0 if otherwise. y_i^* is the latent decision variable, which is a function of the vector z_i of the regressors that affect the participation. The outcome equation represents the continuous outcome variable of hours per month of older workers (h_i), which is observed only when $y_i = 1$. x_i represents the regressors that affect the working hours. γ and β are vectors of unknown parameters, u_i and ε_i are error terms with i.i.d. drawings from a bivariate normal distribution with zero mean, variances σ_u^2 and σ_ε^2, and covariance $\sigma_{u\varepsilon}$.

The mechanism of the model is that when, $y_i^* \leqslant 0$, $y_i = 0$ and h_i cannot be observed, it is defined to be of value 0, and h_i can be observed only when $y_i^* > 0$. The likelihood function of the bivariate sample selection model in the empirical context is given by (see Amemiya [34, p.385])

$$L = \prod_0 Pr(y_i^* \leqslant 0) \prod_1 f_{\varepsilon|u}(h_i \mid y_i^* > 0) Pr(y_i^* > 0), \tag{3.2}$$

where $\prod_0$ stands for the product over those values of i for which $h_i = 0$, $\prod_1$ stands for the product over those value of i for which $h_i = 1$, and $f_{\varepsilon|u}(h_i \mid y_i^* > 0)$ stands for the conditional density of h_i given $y_i^* > 0$.

In the current paper, we will perform the estimation by using two main estimations which are, by making use of the bivariate normal model (BVN), which was first proposed by Heckman [3], and by applying the copula approach to the sample selection model. The likelihood functions in the former case are as followed:

3.2 Standard Heckman's Model (or BVN)

Let the F_u and F_ε be the cumulative distribution functions of the error terms u_i and ε_i, and let both the margins be assumed to be normally distributed: $y^* \sim N(z'\gamma, 1)$ and $h^* \sim N(x'\beta, \sigma_\varepsilon)$, respectively. Then the likelihood function in equation (3.2) can be rewritten as follows (see Amemiya [34, p.386]):

$$L = \prod_0 \Phi(-z'\gamma) \prod_1 \frac{1}{\sigma_\varepsilon} \phi(\frac{h - x'\beta}{\sigma_\varepsilon}) \Phi(\frac{z'\gamma + \theta(h - x'\beta)/\sigma_\varepsilon}{\sqrt{1-\theta^2}}), \tag{3.3}$$

where Φ and ϕ are cdf and pdf, respectively, of the normal distribution, and θ is the dependence parameters.

3.3 Apply Copula Functions to the Sample Selection Model

The estimation of the conditional density function from equation (3.2), however, is complicated. For that reason, Smith [14] applied the copula framework for this model in order to capture it. With that the general form of the likelihood for the sample selection model can be derived as the following (see Smith [14]):

$$L = \prod_0 F_u(0) \prod_1 \{f_\varepsilon(\varepsilon_i) - \frac{\partial}{\partial \varepsilon} F(u_i, \varepsilon_i)\}, \tag{3.4}$$

where $F_u(0) = Pr(y_i^* \leqslant 0)$ and $f_\varepsilon(\varepsilon_i) = \partial(F_\varepsilon)/\partial\varepsilon$,and by using the equation (2.14), we can evaluate the component $\partial F(u_i, \varepsilon_i)/\partial\varepsilon$, which becomes

$$\frac{\partial}{\partial \varepsilon} F(u_i, \varepsilon_i) = \frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)} \times f_\varepsilon, \tag{3.5}$$

Now we substitute equation (3.5) in equation (3.4). Thereafter, the likelihood function can be simplified to

$$L = \prod_0 F_u(0) \prod_1 \{1 - \frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)}\} f_\varepsilon, \tag{3.6}$$

Equation (3.6) is the likelihood function of the copula sample selection model, in which the expression for the component of $(1 - \varphi'(F_\varepsilon)/\varphi'(C_\theta))$ is given by the selected families of copulas, as presented in Table 2.

Table 2: Expressions for $(1-\varphi'(F_\varepsilon)/\varphi'(C_\theta))$

Copula	Expression for $(1-\varphi'(F_\varepsilon)/\varphi'(C_\theta))$
Product copula	$1-F_u$
Gaussian	$1-\Phi\{(\Phi^{-1}(F_u)-\theta\Phi^{-1}(F_\varepsilon))/\sqrt{1-\theta^2})$
AMH	$1-\dfrac{(1-\theta)F_u+\theta F_u^2}{(1-\theta(1-F_u)(1-F_\epsilon))^2}$
FGM	$(1-F_u)(1-\theta F_u(1-2F_\varepsilon))$
Clayton	$1-F_\varepsilon^{-(\theta+1)}(F_u^{-\theta}+F_\varepsilon^{-\theta}-1)^{-\dfrac{1+\theta}{\theta}}$
Frank	$\dfrac{e^{\theta F_u}(e^{\theta F_u}-e^{\theta})}{e^{\theta(F_u+F_\varepsilon)}+e^{\theta}(1-e^{\theta F_u}-e^{\theta F_\varepsilon})}$
Gumbel	$1-C_\theta(F_u,F_\varepsilon)\{(-logF_u)^\theta+(-logF_\varepsilon)^\theta\}^{-1+1/\theta}\{-logF_\varepsilon\}^{\theta-1}F_\varepsilon^{-1}$

Note: 1.F_u denotes $Pr(y_i^*\leqslant 0)$ and F_ε denotes $Pr(y_i^*>0)$.

2.$\overline{F_u}=1-F_u$ and $\overline{F_\varepsilon}=1-F_\varepsilon$

3.This expressions are given as presented in Smith [14]

In the current paper, we apply various copulas to the sample selection model both radially symmetric and asymmetric copulas, such as, the Gaussian, FGM, and Archimedean copulas (AMH, Clayton, Frank, Gumbel and Joe copulas) due to the researcher lacking prior knowledge about the dependence structure. Moreover, we considered different functional forms for margins F_ε and F_u such as normal, logistic, and Student's t distributions. And, at last but not the least, the AIC (Akaike information criterion) and the BIC (Bayesian information criterion) can be used to select between the competing copula models. The AIC and the BIC values are equal to $-2(ln(L)-K)/Q$ and $-2(ln(L)+Kln(Q))/Q$, respectively, where $ln(L)$ is the log–likelihood value at convergence, K is the numbers of parameters, and Q is the number of observations. The better copula model is identified by the lowest values of AIC or BIC.

4 Data

The data set used for this analysis is a sample from the "The Labor Force Survey Whole Kingdom Quarter 3: July– September 2012" conducted by the National Statistical Office. The sample used consists of 2655 observations of older workers, 1512 of whom were employed. This study uses work participation decision (which takes the value 1 if the older worker decides to participate in the labor force, and the value 0 if otherwise) and working hours per month of the older worker (hrpm) as dependent variables for the selection and the outcome equations, respectively. In the selection equation, this study uses the following variables: Education (years of education), Relation (whether the respondent is the head of household), and Marital Status. The regressors of the outcome equation are as follows: Age (age in years), Gender, Education (years of education), and Region.

Table 3: Probit Regression for Selection Equation

Variable	Coefficient	z-Statistic
Constant	0.359***(0.097)	3.707787
Education	0.019***(0.006)	2.971605
Relation	0.469***(0.052)	9.052967
Marital Status	-0.237***(0.036)	-6.656417
McFaddenR^2	0.036350	
S.E.of regression	0.483131	
Log-likelihood	-1748.620	

Note: The dependence variable is work,the symbol "***"
indicates significance at the 1 percent level and SEs are given in parentheses.

Table 4: Least Squares Regression without Controls for Selection Bias

Variable	Coefficient	t-Statistic
Constant	268.895***(21.322)	12.611
Gender	-7.889***(2.936)	-2.687
Age	-0.966***(0.322)	-3.001
Education	-2.176***(0.329)	-6.611
Region	-5.242***(1.225)	-4.279
Adjusted R^2	0.040613	
S.E.of regression	54.84088	
Log-likelihood	-8197.638	
D.W.stat	1.877620	
F-stat	16.99114***	

Note: The dependence variable is work,the symbol "***"
indicates significance at the 1 percent level and SEs are given in parentheses.

5 Results

First, let us estimate the parameters of the marginal distribution models such as the probit model and the least squares regression for the selection and the outcome equations, respectively, which are presented in Table 3 and Table 4. The results of the probit regression, as presented in Table 3, show that whereas education and relation have a significantly positive effect on the labor force participation of older workers, marital status has a negative effect on it. The maximized value of the log–likelihood at convergence is –1748.620. Consider the result of the least square regression without controls for selection bias, as given in Table 4. The results indicate that gender, age, education and region have significantly a negative effect on the working hours per month of the older worker. The maximized value of the log–likelihood at convergence is –8197.638.

Second, in the current paper we consider different functional form for margins F_u and F_ε such as normal, logistic, and Student's t distributions. However, based on the Jarque–Bera test [36] and Shapiro–Wilk test [37]we rejected the null hypothesis that residuals of selection and outcome equations are normally distributed (see Appendix). Moreover, Hasebe [38] have recommended the Student's t distribution as marginal distributions for the outcome equation due to it being the most flexible among the three marginal distributions. Thus, we specify logistic distribution for margins F_u and Student's t distribution for margin F_ε. These ideas differ from the paper of Smith [14] who fixed the margin as a normal distribution.

Third, we considered the sample selection model, which relates to the work participation decision and working hours for older workers in Thailand. We estimated the parameters by using FIML for two main models such as the bivariate normal model (BVN) and by applying the copula approach to the sample selection model such as the Gaussian copula, and the Archimedean families of copulas (AMH, Clayton, Frank, Gumbel and Joe copula). Thus we estimated model with seven different copula functions that were mentioned above and specified the logistic distribution for margins F_u and Student's t distribution for margin F_ε. The results are illustrated in Table 5, where there are 4 good candidate copula based models.

The main result shows that based on AIC and BIC criteria, all of copula based models perform better than the BVN model which restricts the bivariate normality for the joint distribution, especially Frank (L–t) copula. The log–likelihood value at convergence and the BIC value are –9888.156 and 19870.922, respectively (as shown in Table 5). Usually, this copula is radially symmetrical and has strong central dependence, thereby implying that there is no clustering of values in the tail dependence, whether it is left or right tail dependence. The results show that the dependence parameters (θ) which indicates the dependence between the work participation error terms (u_i) and the working hours error term (ε_i) is significantly different from zero in all of the models. This implies that there exists significant dependence between these two error terms, which explains the existence of the selectivity bias. In addition, the parameter υ of the Student–t_υ distribution for Frank (L–t) copula is estimated simultaneously. The results show that the value of υ is about 5. This indicates very thickly tails in the distribution, this guarantee the result of the Jarque–Bera test.

Table 5 shows the similarity of coefficients for all of the models, which is the same as the results obtained in several previous studies (Prieger [35]; Smith [14]; Bhat and Eluru [20]; Genius and Strazzera [10]; Chinnakum, Sriboonchitta and Pastpipatkul [17];Sirisrisakulchai and Sriboonchitta [18]). The selection equation corresponding to the Frank (L–t) copula shows that education and relation have 1 percent significant positive effect on labor force participation of older workers, while marital status has 1 percent significant negative effect on the same. These results indicate that older workers, who are generally the heads of their households, or single or highly educated, are more likely to participate in the labor force. In the outcome equation, all of the variables, namely, age, gender, education, and region were observed to have 1 percent significant negative effect on the working hours per month of the older workers. These results indicate that the older workers

Table 5: Estimates of BVN, Gaussian Copula and the Archimedean Families of Copulas

	standard(N–N)		Guassian(N–N)		Amh(L–t)		Gumbel(L–t)		Frank(L–t)	
Variable	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Selection equation										
Constant	.273***	.088	.273***	.088	.584***	.159	.528***	.161	.565***	.159
Education	.015***	.006	.015***	.006	.029***	.010	.031***	.010	.028***	.010
Relation	.447***	.049	.447***	.049	.775***	.084	.780***	.083	.786***	.083
Marital Status	-.189***	.033	-.189***	.032	-.389***	.060	-.370***	.060	-.383***	.059
Outcome equation										
Constant	246.837***	21.066	246.837***	21.066	259.593***	19.651	257.773***	20.091	255.608***	19.842
Gender	-13.114***	2.976	-13.114***	2.976	-7.621***	2.680	-8.726***	2.917	-8.596***	2.755
Age	-1.117***	.319	-1.117***	.319	-.910***	.298	-.934***	.300	-.920***	.299
Education	-1.499***	.373	-1.499***	.373	-2.335***	.276	-2.175***	.316	-2.220***	.287
Region	-4.943***	1.214	-4.943***	1.214	-4.830***	1.096	-4.879***	1.104	-4.777***	1.099
σ	63.967***	2.364	63.967***	2.364	42.409***	1.732	41.740***	1.648	43.261***	1.924
υ	-	-	-	-	4.547***	.607	4.704***	.623	4.802***	.686
Dependence parameters										
θ	.693***	.056	.693***	.056	.375***	.181	1.202***	.146	1.608***	.668
Kendall'sτ	-	-	.488***	.049	.092***	.050	.168***	.101	.174***	.072
Log-likelihood	-9938.119		-9938.119		-9889.354		-9888.843		-9888.156	
AIC	19898.238		19898.238		19802.708		19801.686		19800.312	
BIC	19962.964		19962.964		19873.318		19872.297		19870.922	

Note: 1)The symbol "***" denotes significance at the 1 percent level, "**" denotes significance at the 5 percent level, and "*" denotes significance at the 10 percent level.
2)N–N denotes normal distribution for margins F_u and F_ε. L–t denotes logistic distribution for margin F_u and Student–t_v distribution for margin F_ε

who are younger, or male, or highly educated, or stay in the Bangkok province are likely to have more working hours per month.

The dependence parameters (θ) between the work participation error terms (u_i) and the working hours error term (ε_i) is positive and highly significant at 1 percent level in this application for the Frank (L–t) copula model, with a corresponding Kendall's τ value of 0.174. This positive dependency implies that unobserved factors which increase (decrease) the propensity of older workers to participate in the labor force also increase (decrease) their working hours. This means that after controlling all the other observed characteristics, the individual who chooses to participate in the labor force will work the higher working hours than an individual who chooses not to participate.

Finally, the copula approach to sample selection, which allows flexibility in the dependence structure, had performed well. This was especially so in the case of the Frank (L–t) copula which provides the best fit. In addition, it performed better than the standard Heckman's method which is restricted to linear dependence. This finding implies that this application is suitable for central dependence, and is not suitable in cases of clustering of values in the tail dependence, whether it is left or right tail dependence. Moreover, the logistic and Student–t_v distribution could be a good distribution for the margins F_u and F_ε in this application respectively.

6 Conclusion

In this paper, we applied the copula approach to a sample selection model to determine the factors affecting the labor force participation and the working hours of older workers in Thailand by using the "The Labor Force Survey of Whole Kingdom Quarter 3: July–September 2012" data set. Importantly, we considered various copulas both radially symmetric and asymmetric copulas, such as, the Gaussian, FGM, and Archimedean copulas (AMH, Clayton, Frank, Gumbel and Joe copulas) due to the researcher not having prior knowledge about the dependence structure. Also, we considered different distributions for margins F_ε and F_u such as normal, logistic, and Student's t distributions.

The main results are the following: first, the copula approach to sample selection model, which allows for flexibility in dependence structure and relax the joint normality assumption, works in the context of the labor force participation and the working hours of older workers in Thailand. Based on the criteria of log–likelihood value, AIC and BIC, the Frank (L–t) copula provides the best fit and performs better than the standard Heckman's method which is restricted to linear dependence. This finding implies that this application is suitable for central dependence, and is not suitable in cases of clustering of values in the tail dependence, whether it is left or right tail dependence. Second, these results show the presence of significant positive dependency of unobserved factors between the two error terms. This implies that selectivity bias exists and that the unobserved factors that increase (decrease) the propensity of older workers to participate in labor force also increase (decrease) the working hours. Third, the results of the coefficient indicate that the concerned policy makers should run campaigns that encourage the various groups of older workers – such as older workers who are not single, who are not heads of households or who have lesser education – to participate in labor force. Furthermore, female workers, or workers who are not single or workers who do not stay in the Bangkok province are likely to register fewer working hours; thus it is imperative that the concerned stakeholders design programs that can increase the working hours of such workers.

7 Appendix

We used the Jarque–Bera normality test and the Shapiro–Wilk normality test for the both residuals of selection and outcome equations. The results of the Jarque–Bera normality test, as given in Table 6, show that the residuals of selection and outcome equations are rejected at 1 percent level of significance. The results of the Shapiro–Wilk test, as given in Table 7, show that the p–value is less than 0.05. Thus we rejected the null hypothesis that residuals of selection and outcome equations are normally distributed.

Table 6: Descriptive statistics

	Residual of selection equation	Residual of outcome equation
Mean	-0.000592	-3.91E-14
Median	0.301140	1.240620
Maximum	0.843356	224.3371
Minimum	-0.754498	-175.8567
Std.Dev.	0.482858	55.27222
Skewness	-0.252447	0.354573
Kurtosis	1.270388	4.711226
Jargue–Bera	359.1415	216.7364
Probability	0.000000	0.000000
Observations	2655	1516

Table 7: The Shapiro–Wilk test for normal data

Variable	Obversations	W	V	Z	p–value
Residual of selection equation	2655	0.82023	275.738	14.440	0.00000
Residual of outcome equation	1516	0.97163	26.139	8.216	0.00000

Acknowledgement(s) : We would like to thank the referees for their comments and suggestions on the manuscript and National Statistical Office for database. The first author was supported by CHE PhD scholarship.

References

[1] National Statistical Office, The Labor Force Survey Whole Kingdom Quarter 1/2013, 2013, `http://web.nso.go.th/en/pub/datapub.`

[2] J.M.Wooldridge, Econometric Analysis of Cross Section and Panel Data, The MIT Press, 2002.

[3] J.J.Heckman,Shadow prices, market wages, and labor supply, Econometrica 42(1974) 679-693.

[4] J.J.Heckman, Sample selection bias as a specification error, Econometrica 47(1)(1979) 153-161.

[5] W.H.Greene, Econometric Analysis, 5th ed., Pearson Education, Inc., Upper Saddle River,New Jersey, 2003.

[6] F.Vella, Estimating models with sample selection bias: a survey, The Journal of Human Resources 33(1998) 127-169.

[7] P.Puhani, The Heckman correction for sample selection and its critique, Journal of Economic Surveys 14 (1)(2000) 53-68.

[8] K.Nawata, Estimation of the female labor supply models by Heckmans two-step estimator and the maximum likelihood estimator, Mathematics and Computers in Simulation 64(2004) 385-392.

[9] K.Nawata, M. Li, Estimation of the labor participation and wage equation model of Japanese married women by the simultaneous maximum likelihood method, J. Japanese Int. Economies 18(2004) 301-315.

[10] M.Genius, E.Strazzera, Applying the copula approach to sample selection modeling,Applied Economics 40(11)(2008) 1443-1455.

[11] L.F.Lee, Some approaches to the correction of selectivity bias, Review of Economic Studies 49(1982) 355-72.

[12] L.F.Lee, Generalized econometric models with selectivity, Econometrica 51(1983) 507-512.

[13] P.K.Trivedi, D.M.Zimmer,Copula modeling: An Introduction for Practitioners,Foundations and $Trends^{®}$in Econometrics 1(1)(2005)1–111.

[14] M.D.Smith, Modelling sample selection using Archimedean copulas, Econometrics Journal,6(2003) 99-123.

[15] B.Eberth, M.D.Smith, Modelling the participation decision and duration of sporting activity in Scotland, Economic Modelling 27(2010) 822-834.

[16] T.Hasebe, W.Vijverberg, A Flexible Sample Selection Model: A GTL-Copula Approach, IZA Discussion Paper No. 7003. 2012

[17] W.Chinnakum, S.Sriboonchitta, and P.Pastpipatkul, Factors affecting economic output in developed countries: A copula approach to sample selection with panel data, International Journal of Approximate Reasoning 54(6)(2013) 809–824.

[18] J. Sirisrisakulchai and S. Sriboonchitta, Factors Affecting Hospital Stay Involving Drunk Driving and Non-Drunk Driving in Phuket, Thailand, In V. N.Huynh, V.Kreinovich, and S.Sriboonchitta, Modeling Dependence in Econometrics in Advances, In Intelligent Systems and Computing, Springer, Heidelberg, vol.251(2014)479–490.

[19] D.Dancer, A.Rammohan, and M. D.Smith, Infant mortality and child nutrition in Bangladesh, Health Economics 17(9)(2008) 1015-1035.

[20] C.R.Bhat, N. Eluru, A Copula-Based Approach to Accommodate Residential Self–Selection Effects in Travel Behavior Modeling, Transportation Research Part B 43(7)(2009) 749–765.

[21] I.N.Sener, C.R.Bhat, A Copula-Based Sample Selection Model of Telecommuting Choice and Frequency, Environment and Planning A. 43(1)(2009) 126–145.

[22] B.Schweizer, Thirty years of copulas, In: G. DallAglio, S. Kotz, and G. Salinetti (eds.): Advances in Probability Distributions with Given Marginals: Beyond the Copulas, The Netherlands: Kluwer Academic Publishers, 1991.

[23] R. B.Nelsen, An Introduction to Copulas, (2nd ed.), Springer-Verlag, New York, 2006.

[24] A.J.Patton, Modelling asymmetric exchange rate dependence, International Economic Review 47(2)(2006) 527-556.

[25] P.Boonyanuphong and S.Sriboonchitta, An Analysis of Volatility and Dependence between Rubber Spot and Futures Prices Using Copula–Extreme Value Theory, In V. N.Huynh, V.Kreinovich, and S.Sriboonchitta, Modeling Dependence in Econometrics in Advances, In Intelligent Systems and Computing, Springer, Heidelberg, vol.251(2014) 431–444.

[26] O.Puarattanaarunkorn and S. Sriboonchitta, Copula Based GARCH Dependence Model of Chinese and Korean Tourist Arrivals to Thailand: Implications for Risk Management, In V. N.Huynh, V.Kreinovich, and S.Sriboonchitta, Modeling Dependence in Econometrics in Advances, In Intelligent Systems and Computing, Springer, Heidelberg, vol.251(2014) 343–366.

[27] S.Sriboonchitta, H. T. Nguyen, A.Wiboonpongse, and J. Liu, Modeling volatility and dependency of agricultural price and production indices of Thailand: Static versus time-varying copulas, International Journal of Approximate Reasoning 54(6)(2013) 793–808.

[28] G.Xue and S.Sriboonchitta, Co–movement of Prices of Energy and Agricultural Commodities in Biofuel Era: A Period–GARCH Copula Approach,In V. N.Huynh, V.Kreinovich, and S.Sriboonchitta, Modeling Dependence in Econometrics in Advances, In Intelligent Systems and Computing, Springer, Heidelberg, vol.251(2014)505–520.

[29] T.Schmidt, Coping with copulas, Forthcoming in Risk Books: Copulas - From Theory to Applications in Finance, Department of Mathematics, University of Leipzig, 2006.

[30] R.B.Nelson, An Introduction to Copulas, Lecture Notes in Statistics, Springer-Verlag, New York, 1999.

[31] M.J.Frank, On the simultaneous associativity of $F(x, y) and x + y_F(x, y)$, Aequationes Mathematicae 19 (1)(1979) 194-226.

[32] D.G.Clayton, A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence, Biometrika 65(1)(1978) 141-151.

[33] H.Joe, Parametric families of multivariate distributions with given marginals, Journal of Multivariate Analysis 46(2)(1993) 262-282.

[34] T.Amemiya, Advanced Econometrics, Cambridge, MA: Harvard, 1985.

[35] J.E.Prieger, A flexible parametric selection model for non-normal data with application to health care usage, Journal of Applied Econometrics 17(2002) 367-392.

[36] C. M.Jarque, A.K.Bera, A test for normality of observations and regression residuals, International Statistical Review 55 (2)(1987) 163-172.

[37] S.S.Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (1965) 591–611.

[38] T.Hasebe, Copula–based maximum-likelihood estimation of sample–selection models, The Stata Journal 13(3)(2013) 547-573.

(Received 30 April 2014)
(Accepted 10 September 2014)