Int. Journal of Economics and Management **8(1):** 104 – 124 (2014) ISSN 1823 - 836X

Fuzzy Modified Error Terms of Sample Selection Model on Married Women Participation in Labour Force

A. YUSRINA[a]*, S. SHAZANI[b] AND L. MUHAMAD SAFIIH[c]

[a,b]*Universiti Malaysia Kelantan*
[c]*Universiti Malaysia Terengganu*

ABSTRACT

Heckman sample selection model (1979) has been widely used in theoretical field and applications field. However, this model involves uncertainties and ambiguities. To solve these problems, a new approach on error terms of Heckman sample selection model has been developed that hybrids with fuzzy concept through triangular fuzzy number. Sample selection model was made up of two equations which were participant equation and wage equation. Since there exist uncertainty in both variables and error terms of the model, therefore fuzzy error terms method was applied on the errors to obtain efficient values which were more reliable in a fuzzy environment. This method was repeated twice on the errors of sample selection model to acquire a strong relationship between the variables and errors which explains uncertainty. The data set were obtained from Malaysian Population and Family Survey 1994 (MPFS 1994). Minimum values of error terms indicates that modified fuzzy sample selection model with fuzzy error terms performs much better when uncertainty and fuzziness exist. Thus, in terms of uncertainty it was found out that the proposed method was more efficient because it explains the data as well as the relationship between the variables in the model much better than its counterpart.

Keywords: Uncertainty, sample selection model, women participation, econometrics, fuzzy number

* Corresponding Author: E-mail: yusrina@umk.edu.my
Any remaining errors or omissions rest solely with the author(s) of this paper.

INTRODUCTION

Heckman sample selection model was introduced by Heckman (1979) in order to manage non-random samples. Individual in a sample study were chosen with fixed criteria and non-random from a population of study. Sample selection model was divided into two parts: structural part and selection part. The structural part was where the samples which were required and this defines the desired criteria. While for the second part, it was the reduced form of the non-random samples taken from the structural part. As mentioned in Froelich (2002), it can improve the attribute of non-random sample and act as a representative for population relationship.

Sample selection model has been widely used in empirical studies, for example in labour force and women wage, education, technology and healthcare (Bhalotra and Sanhueza, 2002; Seshamini and Gray, 2004; Lei, 2005; Madden, 2006). Lewis (1974) discussed on the participation of working women in labour force and selection bias in determining the decision of women to participate in labour force in long term. There were two main reasons why selection bias exist in the model. According to Heckman (1979), observed individual were selected according to a set of given criterias and also due to the action taken by the analyst. For instance, in the study of women labour force, the information of family stabilization were usually needed hence analysis for repeated observations can be done.

Two-step method (Lola *et al*., 2009; Newey, 2009) was a common method used to estimate SSM as the application was much easier to estimate the model value (Vella, 1998; Martins, 2001; Lei, 2005). Sample selection bias which exist in the samples cause inconsistency in the estimation of the model. However, it can be reduced with this method. Inverse Mills ratio (Vella, 1998; Lei, 2005) was the error terms in the first step, probit step which explains the participation while the second step, ordinary least square step was only estimated on the participation of individual in the study. Mroz (1984) introduced Mroz sequence criteria where only married women with complete information were included in the study. While the rest will be removed from the sample study. The samples were reduced because of the reduction of married women samples which were unusable.

Vella (1998) discussed on the different types of estimation in econometric modelling and selection bias which exist in sample selection model. In his paper, there were three types of sample selection model discussed which was parametric, semiparametric and nonparametric model. According to Vella (1998), Heckman model (1979) manage to solve selection bias. Puhani (2000) discussed on selection bias and the effect of variables of education attainment towards the outcome of the study. In this paper, individuals who did not participate in labour force, were removed from the observations.

Martins (2001) discussed on the participation of married women in labour force in Portugal by using crisp parametric and crisp semiparametric sample selection model. Lei (2005) also discussed on both sample selection but on applications of man and women participation of labour force in Canada. Married women participation in labour force were divided into two categories, which were participant and non-participant. Since, there were no outcome values for non-participant of women, this caused selection bias to occur (Martins, 2001).

Solo and Orunsola (2007) also discussed on women wage in Nigeria where they concluded that the number of children determine women decision to participate in labour force by using two step estimation and maximum likelihood estimation. Nevertheless, the research did not discussed on uncertainty. Therefore, fuzzy concept in sample selection model was first discussed in Muhamad Safiih *et.al* (2006,2008) and Lola *et.al* (2009). In the study, parametric and semiparametric sample selection was developed through fuzzy concept and applied on married women participation. Up until now, only their research discussed on the fuzzy application which exist in the model.

Fuzzy set theory was developed by Zadeh (1965) to represent membership functions in fuzzy system. From this theory, fuzzy attributes were shown in quantitative values. Mathematical approach were used to get an approximation value when information obtained were uncertain, incomplete or inaccurate. This concept has brough to the introduction of fuzzy number which were widely used in fuzzy judgment. Fuzzy number was discussed in Dubois and Prade (1980) where it was used when circumstances cannot be explain in exact values. From fuzzy numbers, it has expanded to triangular fuzzy numbers. There are many types of fuzzy numbers which are triangular, trapezium, Gauss and bell-shape, but triangular was the most commonly used (Pedrycz, 1994, Lola *et.al,* 2009). Triangular number has a left and right triangular fuzzy number as a support for fuzzy number (Dubois and Prade, 1980). The membership function in triangular fuzzy number was much simple as it only have three values and can be reduced to two values for a symmetrical case (Kao and Chyu, 2002).

The purpose of this paper was to introduce a modified form of sample selection model based on its fuzzy error terms that can be used to deal with historical data which contained uncertainty. Sample selection model will be modified with fuzzy error terms method as introduced in Kao and Chyu (2002) as well as fuzzy concept. The first part of this paper discussed on sample selection model and error terms while the second part discussed on fuzzy concept. The new proposed method was developed on the third part of this paper. It was applied on real data of married women participation in Malaysia. The last part discussed on its conclusions.

SAMPLE SELECTION MODEL AND ERROR TERMS

Following (Martins, 2001; Lei, 2005), this model consists of two equations which were participant equation and outcome equation and can be define as follows

$$\begin{aligned} z_i^* &= w_i'\gamma + v_i \\ z_i &= 1 \quad \text{if} \quad y_i^* = w_i'\beta + u_i > 0 \\ z_i &= 0 \quad \text{else} \\ y_i &= y_i^* z_i \qquad\qquad (i = 1,\ldots,N) \end{aligned} \tag{2.1}$$

where z_i^* and y_i^* were endogeneous variables, w_i' and x_i' were exogeneous variables and β and γ were vector parameters. Bias and inconsistency exist on β estimation in Equation (2.1). While u_i and v_i were the random disturbances or error terms can be shown in Equation (2.2).

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & 1 \end{pmatrix} \right) \tag{2.2}$$

The error terms of u_i and v_i in both participant and outcome equation were assumed to be normal, independent and identically distributed i.i.d., i.e. $N \sim (u_i, v_i)$. In Equation (2.2), the existence of covariate value contributes to the relationship between other endogenous and exogenous variables in sample selection model. Since this paper focused on the modified error terms in Equation (2.2), therefore it was rearranged with $i = 1,\ldots,N$ as shown in Equation (2.3) and Equation (2.4).

Participant equation error term:

$$\begin{aligned} y_i^* &= x_i'\beta_i + u_i > 0 \\ &= u_i + x_i'\beta_i \\ u_i &= y_i^* - (x_i'\beta_i) \end{aligned} \tag{2.3}$$

Outcome equation error term:

$$\begin{aligned} z_i^* &= w_i'\gamma + v_i \\ v_i &= z_i^* - (w_i'\gamma) \\ z_i &= 1 \qquad \text{if} \qquad y_i^* = x_i'\beta + u_i > 0 \\ z_i &= 0 \qquad \text{else} \\ y_i &= y_i^* z_i \end{aligned} \tag{2.4}$$

where both error terms were highly correlated and identically distributed as shown in Equation (2.5) and Equation (2.6).

$$u_i \sim N(0(\sigma_u^2, 1)) \tag{2.6}$$

The two estimation method widely used in sample selection model were maximum likelihood estimation (Froelich, 2002) and Heckit two step estimation (Heckman, 1979). Although maximum likelihood estimation was more efficient than Heckit's estimation, however the complexity of this method was quite difficult to execute for a simple equation such as sample selection model (Nawata, 2004; Lola *et al.,* 2009). Therefore, Heckman (1979) introduced two step estimation method to overcome this problem. Heckit's estimation was applied in sample selection model as it was more consistent in terms of estimating the parameters. Hence, variable analysis can be easily done in complex model. The two steps were probit which was to estimate the participant equation and least square method to estimate the outcome equation as shown in Equation (2.1).

Probit step was used to estimate the γ value which were obtained from all the observation of Prob $(z_i = 1 \mid w) = E[z \mid w] = \Phi(w'\gamma) = C(w'\gamma)$ using sample $i = 1, \ldots, N$ to estimate $\hat{\gamma}$. The first step includes conditional expectation as shown in Equation (2.7):

$$\begin{aligned} E(y_i \mid x_i) = E(y_i^* \mid x_i, z_i^* > 0) &= E(x_i'\beta \mid x_i, z_i^* > 0) + E(u_i \mid x_i, z_i^* > 0) \\ &= x_i'\beta + E(u_i \mid x_i, w_i'\gamma > -v_i) \\ &= x_i'\beta + \sigma_{uv} / \sigma_v^2 \{\phi(w_i'\gamma) / \Phi(w_i'\gamma)\} \\ &= x_i'\beta + \mu\lambda_i \end{aligned} \tag{2.7}$$

From Equation (2.7), $\lambda(\bullet) = \phi(w_i'\gamma)/\Phi(w_i'\gamma)$ was the inverse Mills ratio while $\phi(\bullet)$ and $\Phi(\bullet)$ were the univariate probability distribution and cumulative distribution function, with μ as the covariance between u_i and v_i. β, γ and μ can be consistently estimated using Heckit's two step (Heckman, 1979). The probit model in Equation (2.7) includes probability distribution function and was assumed as standard normal cumulative function (Horowitz, 2004). However, the second step can be quite complex when the error terms have to be fit with the first estimation (Heckman, 1979; Greene, 1981 and Maddala, 1983). Inverse Mills ratio showed the existence of bias in sample selection model. The covariance, σ_{uv} between u and v was assume as 1 in probit model to identify γ. Since the two equations in Equation (2.1) were seen as two different parts, therefore inverse Mills ratio in probit step was inserted into least square method as it becomes an entity. Hence, according to Muhamad Safiih *et al*. (2011) parametric estimation y_i for n observation can be consistently estimated and shown as Equation (2.8):

$$\begin{aligned} z_i &= w_i'\hat{\gamma} + \sigma_{uv}x_i'\beta + v_i \\ z_i^* &= 1(x_i'\beta + u_i > 0) \end{aligned} \tag{2.8}$$

where γ parameter in Equation (2.8) was estimated using least square method to obtained consistently estimated parameter. Since the error terms were highly correlated with each other, the regression variable z on w for selected samples gave inconsistent estimation. It has been well known that it depends on normal distribution assumption for an estimation to be consistent. On the other hand for identifying purpose, variable x_i must contain at least one variable more than variable w_i (Martins, 2001). Therefore, a new modified sample selection model has been developed to handle inconsistency estimation problem where uncertainty exist.

FUZZY CONCEPT

The elements in a crisp set were determined by the membership function that can be either 0 or 1 which is a binary system. The membership function in a crisp set was limited as it must be execute precisely. However, human perception were not always precise and this applies on the surrounding as well. This causes human perception to change often and contains uncertainty (Zadeh, 1965). Element of uncertainty exist to improve fuzzy mathematic model which were caused by other factors (Ekel, 2002). Since multi criteria model such as sample selection model contains more than one criteria, factor and variable, thus explanation using crisp

model will give inaccurate outcome as it contains uncertainties in each of it. Therefore in a fuzzy environment, a fuzzy approach was more suitable to explained these values (Zadeh, 1965).

Let say x is domain while x is the element of set P. Therefore, a fuzzy set $\tilde{P}$ with the respected membership function can be defined as Equation (3.1).

$$\begin{aligned} & \mu_P(x): x \to [0,1] \\ \text{where} \quad & \mu_P(x) = 1 \quad \text{if all } x \text{ is in } P, \\ & \mu_P(x) = 0 \quad \text{if none of } x \text{ is in } P, \\ & 0 < \mu_P(x) < 1 \quad \text{if some of } x \text{ is in } P. \end{aligned} \tag{3.1}$$

Fuzzy numbers exist in the error terms of the new hybrid model. The basic properties of the membership function of fuzzy numbers were as follows:

(a) A fuzzy number of a fuzzy set is concave and normal

(b) Alpha cut for each fuzzy number is in a close interval of real number confidence level.

(c) Each real number in an open interval (a, d) are real numbers.

Fuzzy number which was triangular fuzzy number was applied on the variables and the error terms of sample selection model. The basis of triangular fuzzy numbers were as Equation (3.2):

$$P(x) = \begin{cases} f(x) & \text{for } x \in [a,b] \\ 1 & \text{for } x \in [b,c] \\ g(x) & \text{for } x \in [c,d] \\ 0 & \text{for } x < a \quad \text{and} \quad x > d \end{cases} \tag{3.2}$$

with $a \le b \le c \le d$. f is a continuous function which increases monotonically from point b to 1 while g is a continuous function which decreases from point c to 1. While the membership functions for triangular fuzzy numbers with $\tilde{P}$ were (Kaufmann and Gupta, 1991):

$$\mu_P(q) = \begin{cases} \alpha(q-b/m-b) & \text{if} \quad q \in [b,m] \\ 1 & \text{if} \quad q = m \\ \alpha(a-q/a-m) & \text{if} \quad q \in [m,a] \\ 0 & \text{otherwise.} \end{cases}$$

Two main assumptions for the membership function of $\mu_P(q)$ were:

1. $\mu_P(q)$ increases monotonically with $\mu_P(q)=0$ and $\lim_{q\to\infty}\mu_P(q)=1$ for $q \le x_2$.
2. $\mu_P(q)$ decreases monotonically with $\mu_P(q)=1$ and $\lim_{q\to\infty}\mu_P(q)=0$ for $q \ge x_2$.

Defuzzification was used to change the fuzzy output to crisp values. According to Saneifard and Asghary (2011), there were seven method of defuzzification but the most commonly used is centroid method. This defuzzification method used crisp values which contain uncertainty as a middle value to obtained fuzzy distribution. This is the best method as it included all values in observation and changes it into one value only. It is also less-complicated and more efficient in obtaining the defuzzified values.

Uncertainty exists on the error terms, endogenous variables and exogenous variables of Heckman sample selection model (Heckman, 1979). There were two types of endogenous variables; the first was crisp which have integer values while the second was fuzzy which contain uncertainty. According to Kao and Chyu (2002), a crisp model becomes a fuzzy model if one of the elements in it contains either vagueness or uncertainty. Therefore, in a fuzzy environment a crisp exogenous variable transformed from crisp to fuzzy as it inherits the vagueness of the model. In addition, it became fuzzy because of the mixture of crisp and fuzzy elements. To overcome these uncertainty problems in sample selection model, fuzzy error terms method and fuzzy concept were the basic for development of modified fuzzy sample selection model based on its error terms.

Fuzzy Error Terms of Sample Selection Model

As mentioned before, crisp sample selection model have two error terms (u_i, v_i) with each of it was limited and restricted due to uncertainty and can be improved through fuzzy set. The membership function of fuzzy set was hybrid into the model, to obtained modified fuzzy error terms. It was also known that both crisp error terms were highly correlated, normal and independently identically distributed. Thus, by hereditary the fuzzy error terms also have the same attribute. From Equation (2.3) and Equation (2.4), the error terms which contains uncertainty were hybrid with the properties of fuzzy set in Equation (3.1), thus a modified fuzzy error terms were shown as Equation (4.1) and (4.2).

Participant equation fuzzy error term:

$$\begin{aligned} y_i^* &= x_i'\beta + \tilde{u}_i > 0 \\ &= \tilde{u}_i + x_i'\beta_i \\ \tilde{u}_i &= y_i^* - (x_i'\beta_i) \end{aligned} \tag{4.1}$$

Outcome equation fuzzy error term:

$$\begin{aligned} & z_i^* = w_i'\gamma + \tilde{v}_i \\ & \tilde{v}_i = z_i^* - (w_i'\gamma) \\ & z_i = 1 \qquad \text{if} \qquad y_i^* = x_i'\beta + \tilde{u}_i > 0 \\ & z_i = 0 \qquad \text{else} \\ & y_i = y_i^* z_i \end{aligned} \tag{4.2}$$

where z_i^* and y_i^* were endogeneous variables, w_i' and x_i' were exogeneous variables and β and γ were vector parameters. By Equation (4.1) and Equation (4.2), fuzzy error terms distribution of $\tilde{u}_i$ and $\tilde{v}_i$ can be obtained as in Equation (4.3) and Equation (4.4).

$$\tilde{u}_i \sim N(0(\sigma_u^2, 1)) \tag{4.3}$$

$$\tilde{v}_i \sim N(0(\sigma_v^2, 1)) \tag{4.4}$$

From Equation (4.1) and Equation (4.2), the variables were fuzzy as well as it inherited the attribute of the model which contains uncertainty. Therefore, Equation (4.1) which was hybrid with fuzzy error terms method (Kao and Chyu, 2002) were shown in Equation (4.5). The steps shown were according to Kao and Chyu (2002).

$$\tilde{u}_i = \tilde{y}_i^* - (\tilde{x}_i\beta_i) \tag{4.5}$$

with $\tilde{x}_i'$ and $\tilde{y}_i^*$ as fuzzy variables and β_i as parameter. Let say $R = (-b, 0, a)$ was the estimation of $\tilde{u}_i$ with b is left or lower value and a is the right or upper value. Thus, the estimation variable became

$$\hat{\tilde{y}}_i^* = \tilde{x}_i'\beta_i + \tilde{R} \tag{4.6}$$

The fuzzy observation value, $\hat{\tilde{y}}_i^*$ and fuzzy estimation $\hat{\tilde{y}}_i$ in triangular fuzzy number form were $y_i^* = (y_{ib}^*, y_{im}^*, y_{ia}^*)$ and $\hat{y}_i = (\hat{y}_{ib}, \hat{y}_{im}, \hat{y}_{ia})$ respectively. If D_{1i} is the estimation error for u_i in participant equation of Equation (4.6), it can be minimized as

$$\text{Min} \sum_{i=1}^{n} D_{1i}$$

$$\text{s.t.} \qquad D_{1i} = \int_{s_{\tilde{y}_i^*} \cup s_{\hat{y}_i^*}} \left| \mu_{\tilde{y}_i^*}(y) - \mu_{\hat{\tilde{y}}_i^*}(y) \right| dy \tag{4.7}$$

It was known that $\hat{\tilde{y}}_i^* = \tilde{x}_i' \beta_i + \tilde{R}_{1i}$ where $R_{1i} = (-b, 0, a)$. If $b_{\min}$ and $a_{\min}$ as the least left and right value, therefore it can be obtained from $(\hat{y}_{im}^* - \hat{y}_{ib}^*) \geq b_{\min}$, $(\hat{y}_{ia}^* - \hat{y}_{im}^*) \geq a_{\min}$. The estimation variable of $\hat{\tilde{y}}_{ia}^*$ were represented as follows

$$\hat{y}_i^* = (\hat{y}_{ib}^*, \hat{y}_{im}^*, \hat{y}_{ia}^*)$$

with the parameter values respectively were

$$\hat{\tilde{y}}_{ib}^* = (\beta_1 x_{ib}' - b),\ \hat{\tilde{y}}_{im}^* = (\beta_1 x_{im}'),\ \hat{\tilde{y}}_{ia}^* = (\beta_1 x_{ia}' + a)$$

Hence, in simplified form

$$\begin{aligned} \hat{y}_i^* &= (\hat{y}_{ib}^*, \hat{y}_{im}^*, \hat{y}_{ia}^*) \\ &= (\beta_1 x_{ib}' - b, \beta_1 x_{im}', \beta_1 x_{ia}' + a) \\ &= (\beta_0 + \beta_1 x_{ib}' - b, \beta_0 + \beta_1 x_{im}', \beta_0 + \beta_1 x_{ia}' + a) \end{aligned} \tag{4.8}$$

As for the i^{th} observation, the differences of were calculated as Equation (4.9):

$$\begin{aligned} \tilde{D}_{1i} = {} & 0.5(y_{ia} - y_{ib}) + 0.5\left[a + b + \beta_1 (x_{ia}' - x_{ib}')\right] \\ & - 2(0.5)\left[y_{ia} - \beta_0 + \beta_1 (x_{ib}' - b)\right] \alpha_{1i} \end{aligned} \tag{4.9}$$

where

$$\begin{aligned} \alpha_{1i} &= \text{height}(\tilde{y}_i^* \cap \hat{\tilde{y}}_i^*) \\ &= (y_{ia} - \hat{y}_{ib}) / (y_{ia} - y_{im} + \hat{y}_{im} - \hat{y}_{ib}) \end{aligned}$$

as α_{1i} value was the membership function of the $\hat{\tilde{y}}_i^*$ and $\hat{\tilde{y}}_i^*$ intersection. In this study, the alpha cut values were determined based on the experience and observation done by the fuzzy expert. After the first *Di* $\tilde{D}_{1i} = \tilde{u}_i$ was obtained, it was inserted back into Equation (4.5) to obtained Equation (4.10):

$$\tilde{D}_{1i} = \tilde{u}_i = \tilde{y}_i^* - (\tilde{x}_i'\beta_i) \tag{4.10}$$

Centroid method for defuzzification was used the find the crisp values of the fuzzy error terms and variables. If u_{im}, y_{im}^* and x_{im}' were the defuzzify values of $\tilde{u}_i$, $\tilde{y}_i^*$ and $\tilde{x}_i'$, therefore the crisp values were:

$$u_{im} = \int_{-\infty}^{\infty} u\mu_{ui}(u)du \Big/ \int_{-\infty}^{\infty} \mu_{u_i^*}(u)du, \quad y_{im}^* = \int_{-\infty}^{\infty} y\mu_{y_i^*}(y)dy \Big/ \int_{-\infty}^{\infty} \mu_{y_i^*}(y)dy$$

and

$$x_{im}' = \int_{-\infty}^{\infty} \mu_{x_i'}(x)dx \Big/ \int_{-\infty}^{\infty} \mu_{x_i'}(x)dx$$

With all the observations of u_{im}, y_{im}^* and x_{im}' became

$$u_{im} = 1/3(u_{ib} + u_{im} + u_{ia}), \quad y_{im}^* = 1/3(y_{ib}^* + y_{im}^* + y_{ia}^*)$$

and

$$x_{im}' = 1/3(x_{ib}' + x_{im}' + x_{ia}')$$

Since symmetrical fuzzy number was used on this study, therefore the crisp values of the error terms and variables were the middle values which were u_{im}, y_{im}^* and x_{im}' respectively. Through this defuzzification method, Equation (4.10) were rearranged so that crisp values of participation equation were obtained as follows:

$$\begin{aligned} D_{1i} = u_i &= y_i^* - (x_i'\beta_i) \\ u_i &= y_i^* - (x_i'\beta_i) \\ y_i^* &= u_i + x_i'\beta_i \end{aligned} \tag{4.11}$$

The steps from Equation (4.5) until Equation (4.11) were repeated once on the outcome equation in Equation (4.2) to obtained the second estimation error value, $D_{2i} = v_i$. Thus, fuzzy error term method on the second equation was shown in Equation (4.12).

$$\begin{aligned} D_{2i} = v_i &= z_i^* - (w_i'\gamma) \\ v_i &= z_i^* - (w_i'\gamma) \\ z_i^* &= w_i'\gamma + v_i \end{aligned} \tag{4.12}$$

A new modified fuzzy sample selection model with fuzzy error terms has been developed and were shown in Equation (4.13) if else

$$\begin{aligned} &\tilde{z}_1^* = \tilde{w}_i'\gamma + \tilde{v}_i \\ &\tilde{z}_i = 1 \quad \text{if} \quad \tilde{y}_i^* = \tilde{x}_i'\beta + \tilde{u}_i > 0 \\ &z_i = 0 \quad \text{else} \\ &\tilde{y}_i = \tilde{y}_i^* \tilde{z}_i \qquad (\text{i}=1, ..., N) \end{aligned} \tag{4.13}$$

where $\tilde{z}_i^*$ and were fuzzy endogenous variable, $\tilde{w}_i'$ and $\tilde{x}_i'$ were fuzzy exogenous variable, β and γ were the parameters. Endogenous variables were observable while exogenous variable were unobservable. This new modified model were applied on real data set of married women participation in Malaysia.

Data Set of Sample Selection Model Modification

The data set which was used on this study were obtained from the Malaysian Population and Family Survey 1994 (MPFS-94), which provides information on wages, educational attainment, household composition and other socioeconomic characteristics of married women in Malaysia. The survey data were from 1994 and have been provided by National Population and Family Development Board of Malaysia under Ministry of Women, Family and Community Development Malaysia. The original sample data contained 4444 samples but through sequential criteria (Mroz, 1984) it was reduced to 2792 sample data as the rest of the data was either incomplete or have no recorded family income in 1994 (Lola *et al*. 2009). Therefore, 1100 (39.4%) were married women with participants in labour force while 1692 (60.6%) were non participants. The analysis were limited to women married and aged below 60, not in school or retired, husband present in 1994 and husband reported positive earning for 1994 based on selection rules (Martins, 2001) which was applied to create the sample criteria in choosing for participant and non participant of married women in MPFS-94 data set.

$$Z_i^* = \beta_0 + \beta_1 A + \beta_2 A_2 + \beta_3 EDU + \beta_4 CHD + \beta_5 H_W + u_i$$

$$G_p = \gamma_0 + \gamma_1 EDU + \gamma_2 P_{EXP} + \gamma_3 P_{EXP2} + \gamma_4 P_{EXPCHD} + \gamma_5 P_{EXPCHD2} + v_i$$

$$\begin{aligned} &Z_i = 1 \qquad \text{if} \\ &\text{Outcome } (Y) = G_p \\ &\qquad = \gamma_0 + \gamma_1 EDU + \gamma_2 P_{EXP} + \gamma_3 P_{EXP2} + \gamma_4 P_{EXPCHD} \\ &\qquad\quad + \gamma_5 P_{EXPCHD2} + v_i \\ &Z_i = 0 \qquad \text{else} \\ &y_i = G_p Z_i \qquad\qquad i = 1, \ldots, N \end{aligned} \tag{5.1}$$

From Equation (5.1), the endogenous variables in modified fuzzy sample selection model were Z_i^* and G_p, while the exogeneous variables were **AGE** (A, age in a year divided by 10), **AGE2** (A_2, age square divided by 100), **EDU** (years of education), **CHD** (number of children under 18 living in the family) and **HW** (H_W, log of monthly husband's wage). β_i and γ_i were unknown parameters, while u_i and v_i were the error terms of sample selection model for sample data for women. It was also highly correlated, normal, independent and identically distributed with each other as shown in Equation (2.2). For the determination of wages, standard human capital approach was followed except for potential experience which was calculated age-edu-6 instead of actual work experience with Buchinsky (1998) solution was adopted. Since there were no data for real working experience, therefore potential wage was obtained from potential experience, G_p (Muhamad Safiih *et al*. 2008). When participation $= 1_{(Z_i^* > 0)}$, therefore outcome equation becomes $G_n = \chi_1 EXP + \chi_2 EXP^2$ with EXP as real working experience for non-participant.

Sample selection model in Equation (5.1) were rearranged in order to calculate the error terms as stated in Equation (5.2) and (5.3). Following the justification by Kao and Chyu (2002), sample selection model were also classified as fuzzy. Therefore,

Participation equation error terms:

$$\begin{aligned} &\tilde{Z}_i^* = \beta_0 + \beta_1 \tilde{A} + \beta_2 \tilde{A}_2 + \beta_3 EDU + \beta_4 CHD + \beta_5 \tilde{H}_W + \tilde{u}_i \\ &\tilde{u}_i = \tilde{Z}_i^* - (\beta_0 + \beta_1 \tilde{A} + \beta_2 \tilde{A}_2 + \beta_3 EDU + \beta_4 CHD + \beta_5 \tilde{H}_W) \end{aligned} \tag{5.2}$$

Outcome equation error terms:

$$\tilde{G}_p = \gamma_1 EDU + \gamma_2 \tilde{P}_{EXP} + \gamma_3 \tilde{P}_{EXP2} + \gamma_4 \tilde{P}_{EXPCHD} + \gamma_5 \tilde{P}_{EXPCHD2} + \tilde{v}_i$$

$$\tilde{v}_i = \tilde{G}_p - \left(\gamma_0 + \gamma_1 EDU + \gamma_2 \tilde{P}_{EXP} + \gamma_3 \tilde{P}_{EXP2} + \gamma_4 \tilde{P}_{EXPCHD} + \gamma_5 \tilde{P}_{EXPCHD2}\right) \quad (5.3)$$

Thus, the modified sample selection model with fuzzy error terms for sample data of women were shown in Equation (5.4).

$$\begin{aligned}
&\tilde{Z}_i^* = \beta_0 + \beta_1 \tilde{U} + \beta_2 \tilde{U}_2 + \beta_3 EDU + \beta_4 CHD + \beta_5 \tilde{G}_S + D_{1i} \\
&\tilde{G}_p = \gamma_0 + \gamma_1 EDU + \gamma_2 \tilde{P}_{EXP} + \gamma_3 \tilde{P}_{EXP2} + \gamma_4 \tilde{P}_{EXPCHD} \\
&\qquad + \gamma_5 \tilde{P}_{EXPCHD2} + D_{2i} \\
&Z_i = 1 \qquad \text{if} \\
&\text{Outcome } (Y) = \tilde{G}_p \\
&\qquad = \gamma_0 + \gamma_1 EDU + \gamma_2 \tilde{P}_{EXP} + \gamma_3 \tilde{P}_{EXP2} + \gamma_4 \tilde{P}_{EXPCHD} \\
&\qquad\quad + \gamma_5 \tilde{P}_{EXPCHD2} + D_{2i} \\
&Z_i = 0 \qquad \text{else} \\
&y_i = G_p Z_i \qquad\qquad i = 1, \ldots, N
\end{aligned} \quad (5.4)$$

In Equation (5.1), there were two types of variables which contain uncertainty and integer likewise. It is known that the data contain uncertainty, thus triangular fuzzy number is hybrid into all of the exogenous and endogenous variables except for **EDU** and **CHD**, which were crisp. The fuzzy endogenous variables in this study wer age, wage and potential experience. Since there were some parts of sample selection model contains uncertainty, therefore integer values in exogenous variables namely **EDU** and **CHD** were categorize as fuzzy variables (Kao and Chyu, 2002).

There were two fuzzy endogenous variables, first was binary variables in participation equation and log hourly wages (**HW**) in outcome equation. The binary variables were made up of 2 indicators where 1 represented participant and 0 as non-participant. According to Lola *et al*. (2009), non-participant women were self-employed either in family business, farming or a full-time housewife. While as for fuzzy exogenous variables, they were in both participation and outcome equation. **AGE** was in participant equation, **PEXP** was in outcome equation while **EDU** contained in both equation. The main function of **AGE** and **EDU** were to determine human capital and were expected to have negative probability to be hired in labour force (Lola *et al*. 2009).

The summary of variables used in this study and further discussion on the data variables can be referred in Muhamad Safiih *et al.* (2008). Sample size for both crisp and modified sample selection model was $N = 2792$. The crisp sample selection was calculated without any changes made to the data. While the fuzzy sample selection model was modified towards error terms, exogenous and endogenous variables at different alpha cut values. Comparison was made between the values of participation and outcome for coefficient estimation, signification and error terms against participation equation and outcome equation for both models. The empiricial result for both models with the respective standard error (in bracket) was shown in Table 5.1 and Table 5.2.

Table 5.1 Parametric estimation of sample selection model and modified fuzzy sample selection for participant equation

Participation Equation	(i) Sample Selection Model	(ii) Modification of Fuzzy Sample Selection Model			
Variables/ Alpha Cut		α = 0.2	α = 0.4	α = 0.6	α = 0.8
Constant	-1.6214** (0.7122)	-1.6025** (0.7351)	-1.609** (0.7300)	-1.6147** (0.7244)	-1.6186** (0.7185)
AGE	0.5453** (0.4218)	0.5346** (0.4357)	0.5385** (0.4325)	0.5415** (0.4292)	0.5437** (0.4256)
AGE2	-0.0723** (0.0535)	-0.0712** (0.0551)	-0.0716** (0.0547)	-0.0719** (0.0543)	-0.0722** (0.0539)
EDU	0.05232** (0.0092)	0.05255** (0.0094)	0.05247** (0.00937)	0.05241** (0.00932)	0.0523** (0.00926)
CHD	0.01137** (0.02115)	0.01138** (0.02115)	0.011377** (0.02115)	0.011376** (0.02115)	0.011376** (0.02115)
HW	-0.2232** (0.27070)	-0.2131** (0.2819)	-0.2166** (0.2795)	-0.2195** (0.2768)	-0.2216** (0.2738)

** at 5%level of significance

Table 5.1 showed the empirical result obtained from the first step of Heckit's two step estimation, which was probit for participation equation (Vella (1998), Lei (2005) and Martins (2001)). Column (i) was the crisp sample selection model result while column (ii) was the modified sample selection model result at alpha cut values of 0.2, 0.4, 0.6 and 0.8. In this study, the probability of married women participation in labour force was assumed by probit model. It was known that the

age of women participation was a quadratic function, therefore **AGE** and **AGE2** variables were highly significant with which was stated in Al-Qudsi (1996) article. For example, in **AGE** variable every increment in the probability of married women joining the labour force, there would be a steady increment in a small magnitude. In the crisp sample selection model, when the probability in **AGE** variable changes from 0 to 1 it causes a 54.5% ($Pr_{Y_i=1)}$ or 0.545) change in probability value that will lead to married women participation in labour force. As the alpha cut value increases from 0.2 to 0.8, the probability values were 53.5 % ($Pr_{Y_i=1)}$ or 0.535), 53.9% ($Pr_{Y_i=1)}$ or 0.539) 54.2 % ($Pr_{Y_i=1)}$ or 0.542) and 54.4 % ($Pr_{Y_i=1)}$ or 0.544) when uncertainty exists. This shows that as women's age increases to a certain years, they were more likely to participate in labour force.

The probability estimation of **EDU** and **CHD** were both positive and significant in the crisp model. Here, the similar circumstance were also shown in fuzzy model when the uncertainty exist. It is known that education have a positive impact on married women's decision to participate in labour force. From Table 5.1, 1% change of married women with lower education participate in labour force only in a small ratio of 5.23 % ((Pr $Y_i = 1$) or 0.0523) as in married women participant in Nigeria. The participation probability of married women in labour force was more firmed when there exist uncertainty for education which was 5.23 % ((Pr $Y_i = 1$) or 0.0523) and this value was maintain for alpha values 0.2 to 0.8. This outcome was similar as the justification from Solo and Orunsola (2007) findings.

Married women participation in labour force whom have a stable number of children (number of children ≤ 3) has the probability value of 1.1 % ($Pr_{Y_i=1)}$ or 0.011). In the modified model, there was only a small change in the probability for number of children as alpha value increases. This shows that the number of children a married women has, does not quite effect their decision to participate in the labour force. A negative or insufficient of husband's wage in supplying for the family also effects women decision to participate in labour force as mentioned in women's study in Nigeria and Portugal. Although the probability estimation in crisp model for **HW** variable was negative, it was significant as the result study in Lola *et al.* (2009) and this also applies for the modified model.

A 1% probability change in husband wage, the probability of married women participation decreases to 22 % ($Pr_{Y_i=1)}$ or 0.22). While for the modified model, they were 21.3 % ($Pr_{Y_i=1)}$ or 0.213), 21.7 % ($Pr_{Y_i=1)}$ or 0.217), 21.9 % ($Pr_{Y_i=1)}$ or 0.219) and 22.2 % ($Pr_{Y_i=1)}$ or 0.222) which shows small changes as alpha increases. The probability values obtained, showed the uncertainty inside the model which were not shown by the crisp model.

It was found out that, all the fuzzy variables were consistent although the estimation value increases from 0.2 to 0.8. However, it was still approaching to the estimation probability of the crisp model which shows minimum values of the model. Thus, probit step of participation equation shows that the modified sample selection model were more efficient in dealing with uncertainties.

Table 5.2 Parametric estimation of sample selection model and modified sample selection model for wage equation.

Wage Equation	(i) Sample Selection Model	(ii) Modification of Fuzzy Sample Selection Model			
Variables/Alpha Cut		α = 0.2	α = 0.4	α = 0.6	α = 0.8
Constant	2.2703** (0.1039)	0.5876** (0.0210)	0.3067** (0.04169)	0.0258** (0.0623)	0.0068** (0.0829)
EDU	-0.0020** (0.0019)	-0.0096** (0.0004)	-0.01045** (0.0007)	-0.01125** (0.0011)	-0.01202** (0.0015)
PEXP	0.0584** (0.0574)	0.2421** (0.0117)	0.2764** (0.0231)	0.3103** (0.0344)	0.3439** (0.0458)
PEXP2	-2.4683** (1.3095)	-9.5160** (0.2702)	-10.68** (0.5296)	-11.8418** (0.7885)	-13.0134** (1.0461)
PEXPCHD	-0.0346** (0.0117)	-0.1459** (0.0025)	-0.1655** (0.0048)	-0.1849** (0.0070)	-0.2044** (0.0093)
PEXPCHD2	0.0071** (0.0037)	0.0383** (0.0008)	0.0436** (0.0015)	0.0488** (0.0022)	0.05416** (0.0029)

** at 5%level of significance

Table 5.2 shows the empirical result of least square method given by the second estimation of outcome equation. In column (i) was the crisp sample selection model result while column (ii) was the modified sample selection model result at alpha cut values of 0.2, 0.4, 0.6 and 0.8. The coefficient values obtained were significant with positive values of **PEXP** and **PEXPCHD2** variables, while the others were negative. Married women with working experiences have changes in their income/ wage of 5.84% (0.0584 x 100%) and can reached up to 34.39% (0.3439 x 100%) with each changes of 1% in PEXP when there was uncertainty in the model.

The least square results show that married women income increases depending on their working experiences as stated in Phimister (2004). This finding was clarified with the increment of women's income as alpha approaches 1 in the modified model. The results for potential experience (PEXP and PEXP2) were significant as the findings obtained from Martins (2001). Although negative coefficients were obtained for PEXPCHD and PEXPCHD2 variables, it was in line with the economic

theory as stated in Coelho *et al*. (2005). It was also found out that education does have a negative relationship with women's income which was the same as the study in Nigeria. For each 1% increment of education of married women, it reduces their income to 0.20% (0.0020 x 100%). In fuzzy environment, women's income decline between 0.97% (0.0097 x 100%) up till 1.2% (0.0120 x 100%) with every 1% of change in education attainment which supports the value obtained in the crisp model. This situation occurs probably due to the lack of suitable occupation in certain parts of Malaysia, such as rural areas. It applies to those who have higher education which was also discussed in the study of married women in Nigeria.

The modification sample selection model result for wage equation shows that the error terms in the fuzzy coefficient were much smaller than the error terms in the crisp coefficient. In the modified model, there were only small changes found on the coefficient estimation as the alpha cut values increases from 0.2 to 0.8. Although the fuzzy coefficient values were spreading from crisp coefficient, nevertheless the values of the error terms was smaller in the modified model compared to the crisp model. This shows a strong relationship between the fuzzy variables in sample selection model when there was uncertainty. Null hypothesis test were also done with zero correlation ($\rho = 0$) at 5% significant level in both crisp and modified sample selection model. In the participation equation, family size and husband's wage were failed to be rejected at 5% significant level which shows that both factor rely to each other in women's decision to participate in labour force. These findings were similar to the study of married women in Canada, Portugal and Nigeria. While as for the wage equation, all **PEXP**, **PEXP2**, **PEXPCHD** and **PEXPCHD2** variables were also failed to be rejected at 5% significant level with smaller value of ρ (less than 0.05) which was similar to Phimister (2004) study. This shows that the variables with minimize error terms in the wage equation were significant towards women's income as stated in previous study. Goodness fit test, R^2 were also done on both crisp and modified model, which shows the existence of bias in the model. The smallest R^2 value was obtained in the crisp model. While as for the modified model, R^2 value increases as alpha cut value increases from 0.2 to 0.8. This shows that the sample data of married women used were compatible with the modified model which was similar to the findings in married women study in Nigeria.

The sample selection model only shows the crisp part. However, the fuzzy variables in the modified sample selection model which contains uncertainty performs much better and were much reliable than the crisp model. Furthermore, the results for modified model proves to be more efficient, significant and wholly in explaining fuzzy and vagueness. Thus, the modified model was the best alternative in explaining uncertainties that exist in a model.

CONCLUSION

From the findings, it can be concluded that uncertainty can be explained more efficiently by using fuzzy approach. A new modified model was developed to explained sample selection model which has suffered previously from the deficiency through the use of crisp method. In this study, uncertainties were explained by minimizing the error terms as well as variables of the sample selection model. Since the sample data of married women used were historical data, crisp method were not suitable to explained these data as it the error terms, endogenous and exogenous variables contains uncertainty. Fuzzy variables which were used in sample selection model becomes more efficient as uncertainty and vagueness can be explained through these fuzzy methods. The whole observation of married women participation in labour force can be seen more throughly compared to the crisp approach. Thus, through modified fuzzy sample selection model, uncertainty can be reduced and sample selection model performs more efficiently where there exist vagueness. Not only modification of sample selection model successfully explained uncertainty, but it also manage to give a significant contribution towards selection models and econometric models. The minimize error terms causes the expert to understand the relationship between the variables in the sample data much better. Finally, since the results obtained were more accurate, therefore it can help the economy model legislation to improve the new policy especially for married women wage in Malaysia in the future.

REFERENCES

Al-Qudsi, S.S. (1996) Labor participation of Arab women: estimates of the fertility to labor supply link, *Economic Research Forum*.

Bhalotra, S. & Sanhueza, C. (2002) Parametric and semi-parametric estimations of the return to schooling in South Africa, Oxford University.

Buchinsky, M. (1998) The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics*, 13: 1-30.

Coelho, D., Veiga, H. & Veszteg, R. (2005) Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal, UFAE and IAE working papers 636.05, Unitat de Fonaments de l'Analisi Economica (UAB) and Institut d'Analisis Economica (CSIC).

Dubois D. & Prade, H. (1980) Fuzzy sets and systems: Theory and applications, *Academic Press*, New York.

Ekel, P. Y. A. (2002) Fuzzy sets and models of decision making, *Computers and Mathematics with Applications* 44: 863-875.

Froelich, M. (2002) Semiparametric estimation of selectivity models, Nova Science Publishers, New York.

Greene, W.H. (1981) Sample selection bias as a specification error: comment. *Econometrica* 49(3): 795-798.

Heckman, J. (1979) Sample selection bias as specification error, *Econometrica*, 47: 153-161.

Horowitz, J.L. (2004) Semiparametric models. In *Handbook of computational statistics*, ed. J.E. Gentle, W. Hardle, & Y. Mori. Springer: New York.

Kao, C. & Chyu C.L. (2002) A fuzzy linear regression model with better explanatory power. *Fuzzy Sets and Systems* 126: 401-409.

Kaufmann, A. & Gupta, M. (1991) Introduction to fuzzy arithmetic-theory and applications, Van Nostrand Reinhold, New York.

Lei, J. J. (2005) Parametrics and semiparametric estimations of the return to schooling of wage workers in Canada, Simon Fraser University.

Lewis, H. G. (1974) Comments on selectivity biases in wage comparisons. Journal of Political Economy 82: 1145-1156.

Lola, M. S., Kamil, A. A. & Abu Osman, M. T. (2009) Fuzzy parametric of sample selection model using Heckman two step estimation models, *American Journal of Applied Sciences* 6(10): 1845-1853.

Madden, D. (2006) Sample selection versus two-part models revisited: the case of female smoking and drinking, Health, Economics and Data Group, Working paper 06/12, University College Dublin.

Maddala, G. S. (1983) Limited-dependent and qualitative variables in econometrics, Vol 1 USA: Cambridge University Press.

Martins, M.F.O. (2001) Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in Portugal, *Journal of Applied Econometrics* 16:23-29.

Mazumdar, D. (1991) The urban labor market and income distribution: A study of Malaysian, Oxford University Press, New York, 1-375.

Mroz, T.A. (1984) The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Ph.D. dissertation, Stanford University London, UK.

Muhamad Safiih, L., Basah Kamil, A. A. & Abu Osman, M. T. (2006) Fuzzy approach to sample selection model, *WSEAS Transactions on Mathematics*, Issue 6(5): 706-712.

Muhamad Safiih, L., Basah Kamil, A. A. & Abu Osman M. T. (2008) Fuzzy semi-parametric sample selection model case study for participation of married women, *WSEAS Transactions on Mathematics Issue 3*(7):112-119.

Muhamad Safiih, L. (2011) Fuzzy parametric and fuzzy semiparametric sample selection model, Thesis submitted for the Degree of Philosophy (Science), University Sains Malaysia, Malaysia.

Nawata, K. (2004) Estimation of the female labor supply models by Heckman's two-step estimator and the maximum likelihood estimator, *Mathematics and Computers in Simulation* 64: 385-292.

Newey, W. K. (2009) Two-step series estimation of sample selection models. The Econometrics Journal, 12(s1), S217-S22.

Pedrycz, W. (1994) Why triangular membership functions? *Fuzzy Sets and Systems* 64: 21-30.

Phimister, E. (2004) Urban effects on participation and wages: are there gender differences? Centre for European Labour Market Research, Department of Economics, University of Aberdeen, Discussion paper 2004-04.

Puhani, P. A. (2000) The Heckman correction for sample selection and its critique, *Journal of Economic Surveys* 14: 53-68.

Saneifard, R. & Asghary, A. (2001) A method for defuzzification based on probability density function (II), *Applied Mathematical Science* Vol 5(28): 1357- 1365.

Seshamini, M. & Gray, A. (2004) Ageing and health care expenditure: The red herring argument revisited, *Health Economics* Vol 13: 303-314.

Solo, O. & Orunsola, E. O. (2007) Sample selection analysis of wage equation for women in Ondo state Nigeria, *Journal of Social Sciences* 3(3):127-133.

Vella, F. (1998) Estimating models with sample selection bias: A survey, *The Journal of Human Resources*; 33: 127-169.

Zadeh, L.A. (1965) Fuzzy sets, *Information and Control 8*: 338-353.