Fuzzy approach to Semi-parametric sample selection model

L. MUHAMAD SAFIIH [1]*, A.A.BASAH KAMIL[2], M. T. ABU OSMAN [3]
[1, 3]Mathematics Department
Faculty of Science and Technology, University Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, MALAYSIA.

[2]School of Distance Learning
Universiti Sains Malaysia
11800 USM Penang, MALAYSIA.

Abstract: The sample selection model studied in the context of semi-parametric methods. With the deficiency of the parametric model, such as inconsistent estimators etc, the semi-parametric estimation methods provide the best alternative to handle this deficiency. Semi-parametric of a sample selection model is an econometric model has been found interesting application in empirical studies. The issue of uncertainty and ambiguity still become are major problem and complicated in the modeling of semi-parametric sample selection model as well as its parametric. In this study, we will focus in the context of fuzzy concept as a hybrid to the semi-parametric sample selection model. The best approach of accounting for uncertainty and ambiguity is to take advantage of the tools provided by the theory of fuzzy sets. It seems particularly appropriate for modeling vague concepts. Fuzzy sets theory and its properties through the concept of fuzzy number provide an ideal framework in order to solve the problem of uncertainty data. In this paper, we introduce a fuzzy membership function for solving uncertainty data of a semi-parametric sample selection model.

Key-Words:- uncertainty, semi-parametric sample selection model, crisp data, fuzzy number, membership function

1. Introduction

The sample selection model studied in the context of semi-parametric methods. With the deficiency of the parametric model, such as inconsistent estimators etc, the semi-parametric estimation methods provide an alternative to handle this deficiency. The study of semi-parametric econometrics of the sample selection models has received considerable attention from statisticians as well as econometricians in the late of 21st century (see Schafgans, 1996). The termed ñsemi-parametric,ò used as a hybrid model for the selection models, which do not involve parametric forms on error distributions; hence, only the regression function part of the model of interest is used. Consideration based on two perspectives, firstly; no restriction of estimation of the parameters of interest for the distribution function of the error terms, secondly; restricting the functional form of heteroskedasticity to lie in a finite-dimensional parametric family (Schafgans,1996).

Cosslett (1990) considered semi-parametric estimation of two-stage method similar to Heckman (1976) for the bivariate normal case where the first stage consisted of semi-parametric estimation of binary selection model and the second stage consisted of estimating the regression equation. Ichimura and Lee (1990) proposed an extension of applicability of a semi-parametric approach. It was proven that all models can be represented in the context of multiple index frameworks (Stoker, 1986) and shown that it can be estimated by the semi-

parametric least squares method if identification conditions are met (see also, Klein and Spady (1993), Gerfin (1996), Martins (2001), Khan and Powell (2001)). Frankly speaking, the previous study in this area concentrates on sample selection model used parametric, semi-parametric or nonparametric approaches. More specifically, none of these researchers put efforts into studies that analyzed semi-parametric sample selection models in the context of fuzzy environment like fuzzy sets, fuzzy logic or fuzzy sets and systems (M.Safiih (2007)).

The purpose this paper is to introduce a membership function of a sample selection model in which historical data contains some uncertainty. With this, provides an ideal framework to deal with problems in which there does not exist a definite criterion for discovering what elements belongs or do not belongs to a given set. Fuzzy set defines by a fuzzy sets in a universe of discourse U is characterized by a membership function denoted by the function μ_A maps all elements of U that take the values in the interval [0,1] that is $A: X \to [0,1]$ (Zadeh, 1965). The concept of fuzzy sets by Zadeh is extended from the crisp sets, that is the two-valued evaluation of 0 or 1, {0, 1}, to the infinite number of values from 0 to 1, [0, 1]. (see Terano *et.al*. 1994).

2 Representation of uncertainty

Generally, fuzzy number represents an approximation of some value which is in the intervals terms $[c^{(l)}, d^{(l)}]$, $c^{(l)} \le d^{(l)}$ for l 0,1,..,n, is given by the α - cuts at the α -levels μ_l with $\mu_l = \mu_{l-1} + \Delta\mu$, $\mu_0 = 0$ and $\mu_n = 1$, usually provide a better job set to compare the corresponding crisp values. As widely practiced used, each α-cuts ${}^{\alpha}A$ of fuzzy set A are closed and related with interval of real numbers of fuzzy numbers for all $\alpha \in (0,1]$ and based on the coefficient $A(x)$: if ${}^{\alpha}A \ge \alpha$ then ${}^{\alpha}A = 1$ and if ${}^{\alpha}A < \alpha$ then ${}^{\alpha}A = 0$ which is the crisp set ${}^{\alpha}A$ depends on α.

Closely related with a fuzzy number is the concept of membership function. In this concept, the element of a real continuous number in the interval [0,1] or in other word representing partial belonging or degree of membership are used. The triangular membership function is used. These represented as a special form as:

$$\mu_A(x) = \begin{cases} \dfrac{(x-c)}{(n-c)} & if\ x \in [c,n] \\ 1 & if\ x = n \\ \dfrac{(d-x)}{(d-n)} & if\ x \in [n,d] \\ 0 & otherwise \end{cases} \quad (1)$$

From that function, the α -cuts of a triangular fuzzy number can be define as a set of closed intervals as

$$[(n-c)\alpha + c, (n-d)\alpha + n], \forall \alpha \in (0,1] \quad (2)$$

and the graph of a typical membership function is illustrated in Figure 1

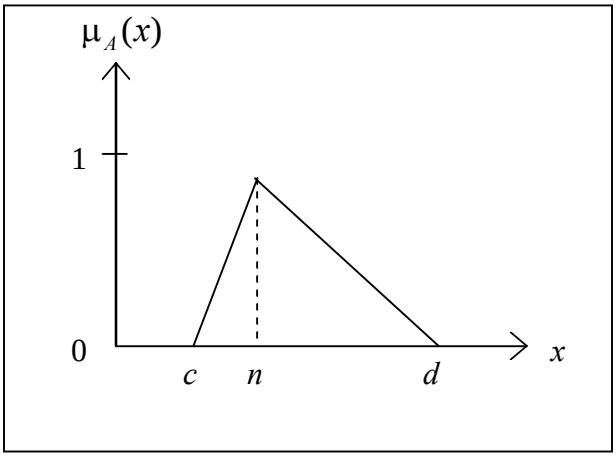

Figure 1: *A triangular fuzzy number*

For the membership function $\mu_A(x)$, the assumptions are as follows:

(i) monotonically increasing function for membership function $\mu_A(x)$ with $\mu_A(x) = 0$ and $\lim_{x\to\infty} \mu_A(x) = 1$ for $x \le n$

(ii) monotonically decreasing function for membership function $\mu_A(x)$ with $\mu_A(x) = 1$ and $\lim_{x\to\infty} \mu_A(x) = 0$ for $x \ge n$

3 The α-cuts representation of fuzzy number

Before going deeper into fuzzy modeling of PSSM, an overview some definitions are presented (Yen *et.at*. (1999), Chen and Wang (1999)) used in this study is related to the existence fuzzy set theory introduced by Zadehôs(1965). The definitions and properties are as follows:

Definition 1*:* the fuzzy function is defined by $f: X \times \tilde{A} \to \tilde{Y}; \tilde{Y} = f(x, \tilde{A})$, where

1) $x \in X$; X is a crisp set;
2) $\tilde{A}$ is a fuzzy set, and
3) $\tilde{Y}$ is the codomain of x associated with the fuzzy set $\tilde{A}$

Definition 2: Let $A \in F(\Re)$ be called a fuzzy number if:

1) exist $x \in \Re$ such that $\mu_A(x) = 1$
2) for any $\alpha \in [0,1]$

$A_\alpha = [x, \mu_{A_\alpha}(x) \geq a]$, is a closed interval with $F(\Re)$ represents all fuzzy sets, $\Re$ is the set of real numbers.

Definition 3: define a fuzzy number A on $\Re$ to be a triangular fuzzy number if its membership function $\mu_A(x): \Re \to [0,1]$ is equal to

$$\mu_A(x) = \begin{cases} \dfrac{(x-l)}{(m-l)} & if\ x \in [l,m] \\ 1 & if\ x = m \\ \dfrac{(u-x)}{(u-m)} & if\ x \in [m,u] \\ 0 & otherwise \end{cases}$$

where $l \leq m \leq u$, x is a model value with l and u be a lower and upper bound of the support of A respectively. Then the triangular fuzzy number denoted by (l, m, u). The support of A is the set elements $\{x \in \Re \mid l < m < u\}$. A non-fuzzy number by convention occurred when $l = m = u$.

Theorem 1*:* The values of estimator coefficients of the participation and structural equations for fuzzy data converge to the values of estimator coefficients of the participation and structural equations for non-fuzzy data respectively whenever the value of $\alpha - \text{cut}$ tend to 1 from below.

Proof. From the centroid method that followed to get the crisp value, the fuzzy number for all observation of w_i as

$$W_{ic} = \frac{1}{3}\left(Lb(w_i) + w_i + Ub(w_i)\right)$$

when there is no $\alpha - cut$. The lower bound and upper bound for each observation referred to by the definition 3 above.

Since we follow the triangular membership function, is followed see Figure 4.2, then $A = \left(Lb(w_{i(\alpha)}), \alpha\right)$ and $B = \left(Ub(w_{i(\alpha)}), \alpha\right)$

where

$$Lb(w_{i(\alpha)}) = Lb(w_i) + \alpha\left(w_i - Lb(w_i)\right)$$

and

$$Ub(w_{i(\alpha)}) = Ub(w_i) + \alpha\left(w_i - Ub(w_i)\right)$$

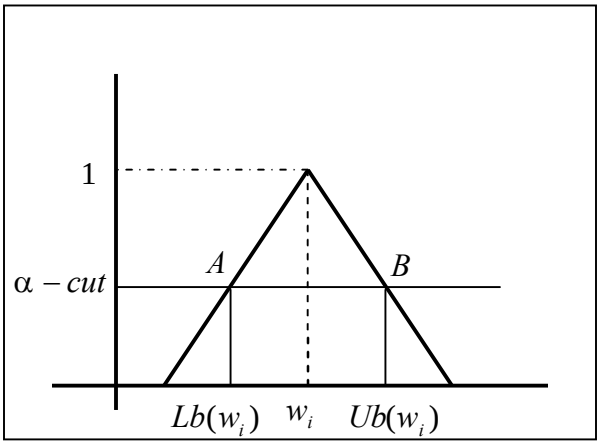

Figure 2: *Membership function and* α *-cut*

Applying the $\alpha - cut$ into the triangular membership function, the fuzzy number is obtained that depends on the given value of the $\alpha - cut$ over the range 0 and 1 is as follow:

$$W_{ic(\alpha)} = \frac{Lb(w_i)+\alpha(w_i - Lb(w_i))+w_i+Ub(w_i)+\alpha(w_i - Ub(w_i))}{3}$$
$$= \frac{Lb(w_{i(\alpha)})+w_i+Ub(w_{i(\alpha)})}{3}.$$

When α approaches 1 from below then $Lb(w_{i(\alpha)}) \to w_i$ and $Ub(w_{i(\alpha)}) \to w_i$. Further obtained is $W_{ic(\alpha)} \to \frac{w_i + w_i + w_i}{3} = w_i$,

$W_{ic(\alpha)} \to w_i$.

The last equation states that when α approaches 1 from below then $W_{ic(\alpha)} \to w_i$. Similarly, for all observations x_i and z_i, $X_{ic(\alpha)} \to x_i$ and $Z_{ic(\alpha)} \to z_i$ respectively, as α tends to 1 from below. This implies that the values of estimator coefficients of the participation and structural equations for fuzzy data converge to the values of estimator coefficients of the participation and structural equations for non-fuzzy data respectively whenever the value of $\alpha - \text{cut}$ tend to 1 from below. □

Definition 4: LR-type fuzzy number denoted as $\tilde{Y}$ with functions $L(Y) = f_1((\frac{1}{\beta})(Y_C - Y))$ and $R(Y) = f_2((\frac{1}{\gamma})(Y - Y_C))$. $\tilde{Y}$ consist the lower bound (Y_L), center (Y_C) and upper bound (Y_U) and satisfying $L(Y_L) = R(Y_U) = 0(\alpha_{\min})$ and $L(Y_C) = R(Y_C) = 1(\alpha_{\max})$. The size of $\tilde{Y}$ is $Y_U - Y_L$ where $\alpha_{\min}$ and $\alpha_{\max}$ can be any predetermined levels.

4 Development of Fuzzy Semi-parametric of Sample Selection Models

Before constructing a fuzzy SPSSM, first, the sample selection model purposed by Heckman (1976) considered. In SPSSM, it is assumed that the distributional assumption of (ε_i, u_i) is weaker than the distributional assumption of the parametric of sample selection model. Then, the sample selection model is now called a semi-parametric of sample selection model (SPSSM).

In the development of SPSSM modeling using fuzzy concept, as a development of fuzzy PSSM, the basic configuration of fuzzy modeling i.e. involved fuzzification, fuzzy environment and defuzzification (see M.Safiih 2007). For fuzzification stage, an element of real-valued input variables converted in the universe of discourse into value of membership fuzzy set. At this approach, a triangular fuzzy number used over all observations. The α - cut method with increment value of 0.2 started with 0 up to 0.8. This is then applied to the triangular membership function to get a lower and upper bound for each observations (x_i, w_i and z_i^*) which is defined as

$$\tilde{w}_{isp} = (w_{il}, w_{im}, w_{iu}), \tilde{x}_{isp} = (x_{il}, x_{im}, x_{iu}) \text{ and}$$
$$\tilde{z}^*_{i\ sp} = (z_{il}, z_{im}, z_{iu}) \quad (3)$$

In order to solve the model in which occurs uncertainties, fuzzy environment such as fuzzy sets and fuzzy number are more suitable, as the processing of the fuzzified input parameters. Since, it is assumed that some original data contains uncertainty, under the vagueness of the original data, the data will then be considered as fuzzy data. That means, each observation considered has are variation values. The upper bound and lower bound of the observation are commonly chosen depending on the each data structure and experience of the researchers. For a large sized of observation, the upper bound and lower bound of each observation are quite difficult to be obtained.

Based on the fuzzy number, a fuzzy SPSSM is built with the form as:

$$\begin{aligned} &\tilde{z}^{*}_{i_{sp}} = \tilde{w}^{'}_{i_{sp}}\gamma + \tilde{\varepsilon}_{i_{sp}} \qquad i = 1,...,N \\ &d_i = 1 \ \ if \ \ d^{*}_{i} = \tilde{x}^{'}_{i_{sp}}\beta + \tilde{u}_{i_{sp}} > 0, \\ &d_i = 0 \ otherwise \ i = 1,...,N \\ &z_i = z^{*}_{ic_{sp}} d_i \end{aligned} \tag{4}$$

The terms $\tilde{w}_{i_{sp}}$, $\tilde{x}_{i_{sp}}$, $\tilde{z}^{*}_{i_{sp}}$, $\tilde{\varepsilon}_{i_{sp}}$ and $\tilde{u}_{i_{sp}}$ are fuzzy numbers with the membership functions $\mu_{\tilde{W}_{isp}}$, $\mu_{\tilde{X}_{isp}}$, $\mu_{\tilde{Z}_{isp}}$, $\mu_{\tilde{\varepsilon}_{isp}}$ and $\mu_{\tilde{u}_{isp}}$, respectively. Since the distributional assumption for the SPSSM is weak, then for the analysis of the fuzzy SPSSM, it is assumed that the distributional assumption is weak.

To find an estimate for γ and β of the fuzzy parametric of sample selection model, one idea is to defuzzify the fuzzy observations $\tilde{W}_{i_{sp}}$, $\tilde{X}_{i_{sp}}$ and $\tilde{Z}^{*}_{i_{sp}}$. That means, converting this triangular fuzzy membership real-value into a single (crisp) value (or a vector of values) that, in the same sense, is the best representative of the fuzzy sets that will actually be applied. Centroid method or the center of gravity method is used i.e. computes the outputs of the crisp value as the center of area under the curve. Let $W_{ic_{sp}}$, X_{icsp} and $Z^{*}_{ic_{sp}}$ be the defuzzified values of $\tilde{W}_{i_{sp}}$, $\tilde{X}_{i_{sp}}$ and $\tilde{Z}^{*}_{i_{sp}}$ respectively. The calculation of the centroid method for $W_{ic_{sp}}$, $X_{ic_{sp}}$ and $Z^{*}_{ic_{sp}}$ respectively via the following formula:

$$W_{ic_{sp}} = \frac{\int_{-\infty}^{\infty} w\mu_{\tilde{w}_i}(w)dw}{\int_{-\infty}^{\infty} \mu_{\tilde{w}_i}(w)dw} = \frac{1}{3}(W_{i_l} + W_{i_m} + W_{i_u})$$

$$X_{ic_{sp}} = \frac{\int_{-\infty}^{\infty} x\mu_{\tilde{x}_i}(x)dx}{\int_{-\infty}^{\infty} \mu_{\tilde{x}_i}(x)dx} = \frac{1}{3}(X_{i_l} + X_{i_m} + X_{i_u})$$

$$Z^{*}_{ic_{sp}} = \frac{\int_{-\infty}^{\infty} z\mu_{\tilde{z}_i}(z)dz}{\int_{-\infty}^{\infty} \mu_{\tilde{z}_i}(z)dz} = \frac{1}{3}(Z_{i_l} + Z_{i_m} + Z_{i_u}) \tag{5}$$

Then the crisp values for the fuzzy observation are calculated following the centroid formula as stated above. To estimate γ_{sp} and β_{sp} of SPSSM approach, applying the procedure as in Powell, then the parameter is estimated for the fuzzy semi-parametric sample selection model (fuzzy SPSSM). Before getting a real value for the fuzzy SPSSM coefficient estimate, first the coefficient estimate values of γ and β are used as a shadow of reflection to the real one. The value of γ and β are then applied to the parameters of the parametric model to get a real value for the fuzzy SPSSM coefficient estimate of $\gamma_{sp}, \beta_{sp}, \sigma_{\varepsilon_{isp}}, u_{i_{sp}}$. The Powell SPSSM procedure is then executed using the XploRe software.

Executing the Powell (Powell, 1987) procedure by XploRe takes the data as input from the outcome equation (x and y, where x may not contain a vector of ones). The vector *id* containing the estimated for the first-step index $x^{'}_{i_{sp}}\beta$, and the bandwidth vector h where h is the threshold parameter k that is used for estimating the intercept coefficient from the first element. The bandwidth h from the second element is used for estimating the slope coefficients. For fuzzy PSSM, follows the above procedure then another set of crisp values $W_{ic_{sp}}$, $X_{ic_{sp}}$ and $Z_{ic_{sp}}$ is obtained. Applied the α - *cut* values on the triangular membership function of the fuzzy observations $\tilde{W}_{i_{sp}}$, $\tilde{X}_{i_{sp}}$ and $\tilde{Z}_{i_{sp}}$ with the original observation, fuzzy data without α - *cut* and fuzzy data with α - *cut* to estimate the parameters of the fuzzy SPSSM. The same procedure above is applying. The parameters of the fuzzy SPSSM are estimated.

5 Conclusion

Basically, modeling take as an important part in estimating the parameters of economic problems. Differing from other system design, the model itself is generated by a mathematical function. In this paper, a description of the development of the FSPSSM has been presented. For handling the uncertainty which is involved in the original data, fuzzy number together with membership function takes an important part derived from expert knowledge.

References:

Chen, T. and Wang. M.J.J. (1999). Forecasting methods using fuzzy concepts, Fuzzy sets and systems.105. p. 339-352.

Cosslett, S. (1991). Semiparametirc estimation of a regression models with sample selectivity, in W.A. Barnett, J. Powell, and G.E. Tauchen (eds), Nonparametric and semiparametric estimation methods in Econometrics and Statistics, Cambridge University Press, p 175-198.

Gerfin, M. (1996). Parametric and semiparametric estimation of the binary response model of labor market participation. Journal of Applied Econometrics. 11. p. 321-339.

Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimation for such models, Annals of Economic and Social Measurement, 5, p. 475-492.

Ichimura, H and Lee L..F. (1990). Semiparametirc least square estimation of multiple index models: Single equation estimation, in W.A. Barnett, J. Powell, and G.E. Tauchen (eds), Nonparametric and semiparametric estimation methods in Econometrics and Statistics, Cambridge University Press.

Khan, S. and Powell, J. L. (2001). Two-step estimation of semiparametric censored regression models. Journal of Econometrics. 103. p. 73-110.

Klein, R. and Spady,R. (1993). An efficient semiparametric estimator of the binary response model, Econometrica, 61, 2, p.387-423.

L. M. Safiih Lola (2007). Fuzzy Semi-parametric of a Sample Selection Models. Ph.D. dissertation. University Science of Malaysia. Penang. Malaysia.

Martin, M.F.O. (2001). Parametric and semiparametric estimation of sample selection models: An empirical application to the female labor force in Portugal. Journal of Applied Econometrics, 16, p. 23-39.

Powell, J.L. (1987). Semiparametric estimation of bivariate latent variable models. Social Systems Research Institute. University of Wisconsin-Madison, Working paper No.8704.

Schafgans, M. (1996). Semiparametric estimation of a sample selection model: estimation of the intercept; theory and applications, Ph.D. dissertation, Yale University New Haven, USA.

Stoker, T.M. (1986). Consistent estimation of scaled coefficients, Econometrica, 54. p. 1461-81.

Terano, T. Asai, K. Sugeno, M. (1994). Applied fuzzy systems, Cambridge. AP Professional.

Yen, K.K., Ghoshray. S. and Roig. G. (1999). A linear regression model using triangular fuzzy number coefficients. Fuzzy sets and systems. 106. p. 167-177.

Zadeh, L.A. (1965). Fuzzy Sets and systems. In: Fox, J., ed., System Theory. Brooklyn, New York. Polytechnic Press.