Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2,

Semi-Parametric of Sample Selection Model Using Fuzzy Concepts

L. Muhamad Safiih[1*] A. A. Kamil[2] M. T. Abu Osman[3]

[1*]Mathematics Department, Faculty Science and Technology, University Malaysia Terengganu,
21030 Kuala Terengganu.
[1*]Institute of Marine Biotechnology,
University Malaysia Terengganu, 21030 Kuala Terengganu.
[2]School of Distance Education University Science Malaysia,
11800 USM, Penang, Malaysia
[3]Kulliyyah Of Information And Communication Technology,
International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur.

The sample selection model has been studied in the context of semi-parametric methods. With the deficiencies of the parametric model, such as inconsistent estimators, semi-parametric estimation methods provide better alternatives. This article focuses on the context of fuzzy concepts as a hybrid to the semi-parametric sample selection model. The better approach when confronted with uncertainty and ambiguity is to use the tools provided by the theory of fuzzy sets, which are appropriate for modeling vague concepts. A fuzzy membership function for solving uncertainty data of a semi-parametric sample selection model is introduced as a solution to the problem.

Key words: Uncertainty, semi-parametric sample selection model, crisp data, fuzzy sets, membership function.

Introduction

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

Gallant and Nychka (1987) studied these methods in the context of semi-nonparametric maximum likelihood estimation and applied the method to nonlinear regression with the sample selection model. Newey (1988) used series approximation to the selection correction term which considered regression s-pline and power series approximations. Robinson (1988) focused on the simplest setting of multiple regressions with independent observations, and described

L. Muhamad Safiih is statistics / econometrics lecturer in the Faculty of Science and Technology, Mathematics Department and fellow in Institute of Marine Biotechnology, University Malaysia Terengganu. Research interest in econometrics modeling, forecasting, applied statistics and fuzzy sets. Email: safiihmd@umt.edu.my. A. A. Kamil is _an Associate Professor_ in the School of Distance Education, Universiti Sains Malaysia. Email: anton@usm.my. M. T. Abu Osman is _a Professor in Kulliyyah of Information and Communication Technology, International Islamic University Malaysia. Research interest in combinatorial group theory, fuzzy mathematics, topology and analysis and mathematics educations. Email: abuosman@iium.edu.my.

extensions to other econometric models, in particular, seemingly unrelated and nonlinear regressions, simultaneous equations, distribution lags and sample selectivity models.

Cosslett (1991) considered semi-parametric estimation of the two-stage method similar to Heckman (1976) for the bivariate normal case where the first stage consisted of semi-parametric estimation of the 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 shown that all models can be represented in the context of multiple index frameworks (Stoker, 1986) and that it can be estimated by the semi-parametric least squares method if identification conditions are met. Andrews (1991) proposed the establishment of asymptotic series estimators for instant polynomial series, trigonometric series and Gallant's Fourier flexible form estimators, for nonparametric regression models and applied a variety of estimands in the regression model under consideration, including derivatives and integrals of the regression function (see also Klein & Spady, 1993; Gerfin, 1996; Vella, 1998; Martin, 2001; Khan & Powell, 2001; Lee & Vella, 2006).

Previous studies in this area concentrated on the sample selection model and used parametric, semi-parametric or nonparametric approaches. None of the studies conducted analyzed semi-parametric sample selection models in the context of fuzzy environment like fuzzy sets, fuzzy logic or fuzzy sets and systems (L. M. Safiih, 2007).

This article introduces a membership function of a sample selection model that can be used to deal with sample selection model problems in which historical data contains some uncertainty. An ideal framework does not currently exist to address problems in which a definite criterion for discovering what elements belong or do not belong to a given set (Miceli, 1998). A fuzzy set, defined by fuzzy sets in a universe of discourse (*U*) is characterized by a membership function and denoted by the function μ_A, maps all elements of *U* that take the values in the interval [0, 1] that is $A: X \rightarrow [0, 1]$ (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]. Brackets { } are used in crisp to indicates sets, whereas square [ ] brackets and parentheses ( ) are used in fuzzy sets to denote real-number closed intervals and open intervals, respectively (see Terano, et al., 1994).

Semi-Parametric Estimation Model

The study of the semi-parametric estimation model involves and considers the two-step estimation approach. The semi-parametric context is a frequently employed method for sample selection models (Vella, 1998) and is a hybrid between the two sides of the semi-parametric approach (i.e., it combines some advantages of both fully parametric and the completely nonparametric). Thus, parts of the model are parametrically specified, while non-parametric estimation issues are used for the remaining part. As a hybrid, the semi-parametric approach shares the advantages and disadvantages of each, in terms that allow a more general specification of the nuisance parameters. In semi-parametric models, the estimators of the parameters of interest are consistent under a broader range of conditions than for parametric models but more precise (converging to the true values at the square root of the sample size) than their nonparametric counterparts.

For a correctly-specified parametric model, estimators for semi-parametric models are generally less efficient than maximum likelihood estimators yet maintain the sensitivity of misspecification for the structural function or other parametric components of the model. In the semi-parametric approach, the differences arise from the weaker assumption of the error term in contrast to the parametric approach. In this study a two-step semi-parametric approach is considered, which generalizes Heckman's two-step procedure. According to Härdle, et al. (1999), Powell (1987) considered a semi-parametric self-selection model that combined the two equation structure of (2.1) with the following weak assumption about the joint distribution of the error terms. For example, the participation equation of the first step is estimated semi-parametrically by the DWADE estimator (Powell, et al., 1989), while applying the Powell (1987) estimator for the second step of the structural equation.

Representation of Uncertainty

Towards representing uncertainty various approaches can be considered. In this study, the representation of uncertainty identified variables by commonly used approaches, that is, putting a range and a preference function to the desirability of using that particular value within the range. In other words, it is similar to the notion of fuzzy number and membership function which is the function μ_A that takes the values in the interval [0, 1], that is, $A : X \to [0,1]$. For more details about representation of uncertainty, this article concentrates on using fuzzy number and membership function.

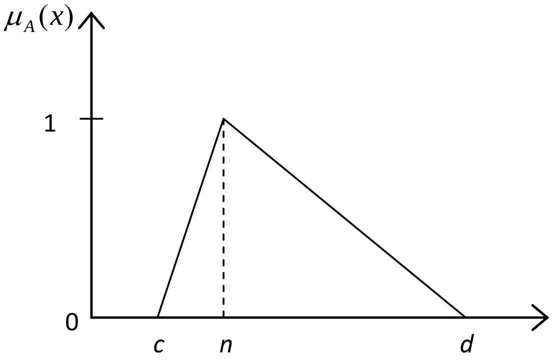
Generally, a fuzzy number represents an approximation of some value which is in the interval terms $[c^{(l)}, d^{(l)}], c^{(l)} \leq d^{(l)}$ for l 0, 1, ..., n, and is given by the α- cuts at the α-levels μ_l with $\mu_l = \mu_{l-1} + \Delta\mu,\quad \mu_0 = 0$ and $\mu_n = 1$. A fuzzy number usually provides a better job set to compare the corresponding crisp values. As widely used in practice, each α-cuts ${}^{\alpha}A$ of fuzzy set A are closed and related with an interval of real numbers of fuzzy numbers for all $\alpha \in (0,1]$ and based on the coefficient $A(x)$: if ${}^{\alpha}A \geq \alpha$ 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 a number representing partial belonging or degree of membership are used. Referring to the definition of the membership function, setting the membership grades is open either subjectively to the researcher, depending on his/her intuition, experience and expertise, or objectively based on the analysis of a set of rules and conditions associated with the input data variables. Here, choosing the membership grades is done subjectively, i.e., reflected by a quantitative phenomenon and can only be described in terms of approximate numbers or intervals such as "around 60," "close to 80," "about 10," "approximately 15," or "nearly 50." However, because of the popularity and ease of representing a fuzzy set by the expert - especially when it comes to the theory and applications - the triangular membership function is chosen. It is called a triangular fuzzy number based on a special type of fuzzy number containing three parameters: the grade starts at zero, rises to a maximum and then declines to zero again as the domain increases with its nature; that is, the membership function increases towards the peak and decreasing away from it, and can be represented as a special form as:

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

The graph of a typical membership function is illustrated in Figure 1.

Figure 1: A Triangular Fuzzy Number



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

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

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 $x \ge n$.

The α-cuts and LR Representation of a Fuzzy Number

Prior to delving into fuzzy modeling of PSSM, an overview and some definitions used in this article are presented (Yen, et al., 1999; Chen & Wang, 1999); the definitions and their properties are related to the existence of fuzzy set theory and were introduced by Zadeh (1965).

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

1. $x \in X$; X is a crisp set, and
2. $\widetilde{A}$ is a fuzzy set, and
3. $\widetilde{Y}$ is the co-domain of x associated with the fuzzy set $\widetilde{A}$.

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

1) $x \in \Re$ such that $\mu_A(x) = 1$,
2) for any $\alpha \in [0,1]$, and
3) $A_\alpha = [x, \mu_{A_\alpha}(x) \ge a]$, is a closed interval with $F(\Re)$ representing all fuzzy sets, $\Re$ is the set of real numbers.

Definition*:* a fuzzy number A on $\Re$ is defined 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 $l \le m \le u$, x is a model value with l and u be a lower and upper bound of the support of A respectively. Thus, the triangular fuzzy number is 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 occurs 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 $\alpha -$cut tends to 1 from below.

Proof:

From the centroid method followed to obtain the crisp value, the fuzzy number for all observation of w_i is

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

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

Because the triangular membership function is followed (see Figure 2) then

$$A = \left(Lb(w_{i(\alpha)}), \alpha\right) \text{ and } B = \left(Ub(w_{i(\alpha)}), \alpha\right),$$

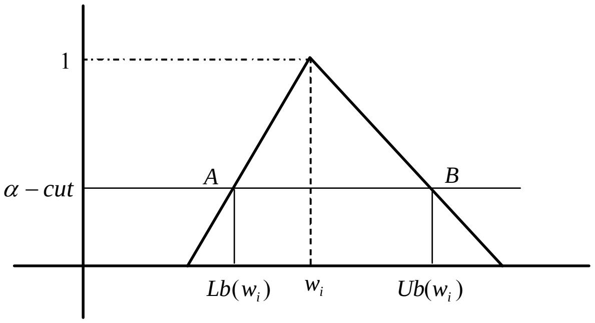
where

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

and

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

Figure 2: Membership Function and $\alpha - \text{cut}$



Applying the $\alpha - cut$ into the triangular membership function, the fuzzy number obtained depending on the given value of the $\alpha - \text{cut}$ over the range 0 and 1 is as follows:

$$W_{ic(\alpha)} = \frac{\begin{array}{c} Lb(w_i) + \alpha\left(w_i - Lb(w_i)\right) \\ + w_i + Ub(w_i) + \alpha\left(w_i - Ub(w_i)\right) \end{array}}{3} = \frac{Lb(w_{i(\alpha)}) + w_i + Ub(w_{i(\alpha)})}{3}.$$

When α approaches 1 from below then $Lb(w_{i(\alpha)}) \to w_i$ and $Ub(w_{i(\alpha)}) \to w_i$, and is obtained as

$$W_{ic(\alpha)} \to \frac{w_i + w_i + w_i}{3} = w_i \ ,$$

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

Thus, 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*:* An LR-type fuzzy number denoted as $\tilde{Y}$ with functions $L(Y) = f_1((\frac{1}{\beta})(Y_C - Y))$ and $R(Y) = f_2((\frac{1}{\gamma})(Y - Y_C))$. $\tilde{Y}$ consists of the lower bound (Y_L), center (Y_C) and upper bound (Y_U). 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.

Theorem 2:

If an LR-type fuzzy number is denoted as $\tilde{Y}'$ with $L(Y')$ and $R(Y')$ functions of $f_1((\frac{1}{k_1\beta})(Y_C^{'} - Y'))$ and $f_2((\frac{1}{k_2\beta})(Y' - Y_C^{'}))$ respectively, then, (Y_L), (Y_C) and (Y_U) of $\tilde{Y}'$ are

$$Y_C^{'} - k_1(Y_C - Y_L), Y_C^{'}$$

and

$$Y_C^{'} + k_2(Y_U - Y_C) .$$

Proof:

Because for $\tilde{Y}$

$$\begin{aligned} L(Y_L) &= f_1\left(\frac{1}{\beta}(Y_C - Y_L)\right) \\ &= R(Y_U) \\ &= f_2\left(\frac{1}{\gamma}(Y_U - Y_C) = 0\right) \end{aligned} \quad ,$$

$$L(Y_C) = f_1(0) = R(Y_C) = f_2(0) = 1 ,$$

then, for $\tilde{Y}'$

$$
\begin{aligned}
L(Y_C^{'} - k_1(Y_C - Y_L)) &= f_1\left(\frac{1}{k_1\beta}(Y_C^{'} - Y_C^{'} + k_1(Y_C - Y_L))\right)\\
&= f_1\left(\frac{1}{\beta}(Y_C - Y_L)\right)\\
&= 0
\end{aligned}
$$

and

$$
\begin{aligned}
R(Y_C^{'} - k_2(Y_U - Y_C)) &= f_2\left(\frac{1}{k_2\gamma}(Y_C^{'} + k_2(Y_U - Y_C) - Y_C^{'})\right)\\
&= f_2\left(\frac{1}{\gamma}(Y_U - Y_C)\right)\\
&= 0
\end{aligned}
$$

$$L(Y_C^{'}) = f_1(0) = R(Y_C^{'}) = f_2(0) = 1$$

Thus, Theorem 2 is proven.

Methodology

Development of Fuzzy Semi-Parametric Sample Selection Models

Prior to constructing a fuzzy SPSSM, the sample selection model purpose by Heckman (1976) is considered. In SPSSM, it is assumed that the distributional assumption of $(\varepsilon_i,\ u_i)$ is weaker than the distributional assumption of the parametric sample selection model. The distributional assumption that exists in Heckman (1979) model is more stringent than anything else. However, the Heckman (1979) estimator becomes inconsistent if the assumption is violated. Härdle, et al. (1999) highlighted that ample reason exists to develop consistent estimators for PSSM with weaker distributional assumptions. Thus, the sample selection model is now called a semi-parametric of sample selection model (SPSSM).

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

$$\tilde{w}_{isp} = (w_{il}, w_{im}, w_{iu}),\ \tilde{x}_{isp} = (x_{il}, x_{im}, x_{iu})$$

and

$$\tilde{z}^*_{isp} = (z_{il}, z_{im}, z_{iu}).$$

Following their memberships functions, respectively defined, results in the following forms:

$$
\mu_{\tilde{w}^*_i sp}(z) = \begin{cases}
\dfrac{(w - w_{il})}{(w_{im} - w_{il})} & if\ \ w \in [w_{im}, w_{im}]\\
1 & if\ \ w = w_{im}\\
\dfrac{(w_{iu} - w_{im})}{(w_{iu} - w_{im})} & if\ \ w \in [w_{im}, w_{iu}]\\
0 & otherwise
\end{cases}
$$

$$
\mu_{\tilde{x}_{isp}}(x) = \begin{cases}
\dfrac{(x - x_{il})}{(x_{im} - x_{il})} & if\ \ x \in [x_{il}, x_{im}]\\
1 & if\ \ x = x_{im}\\
\dfrac{(x_{iu} - x)}{(x_{iu} - x_{im})} & if\ \ x \in [x_{im}, x_{iu}]\\
0 & otherwise
\end{cases}
$$

and

$$
\mu_{\tilde{z}_{isp}}(z) = \begin{cases}
\dfrac{(z - z_{il})}{(z_{im} - z_{il})} & if\ \ z \in [z_{im}, z_{im}]\\
1 & if\ \ z = z_{im}\\
\dfrac{(z_{iu} - z_{im})}{(z_{iu} - z_{im})} & if\ \ z \in [z_{im}, z_{iu}]\\
0 & otherwise
\end{cases}
$$

In order to solve the model in which uncertainties occur, fuzzy environments such as fuzzy sets and fuzzy numbers are more suitable as the processing of the fuzzified input parameters. Because, it is assumed that some original data contains uncertainty, under the vagueness of the original data, the data can be considered as fuzzy data. Thus, each observation considered has variation values. The upper and lower bounds of the observation are commonly chosen based on the data structure and experience of the researchers. For a large-sized observation, the upper and lower bounds of each observation are difficult to obtain.

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

$$\widetilde{z}^{*}_{i_{sp}} = \widetilde{w}^{'}_{i_{sp}} \gamma + \widetilde{\varepsilon}_{i_{sp}} \qquad i = 1,\ldots,N$$

$$d_i = 1 \;\; if \;\; d_i^{*} = \widetilde{x}^{'}_{i_{sp}} \beta + \widetilde{u}_{i_{sp}} > 0,$$

$$d_i = 0 \; otherwise \; i = 1,\ldots,N$$

$$z_i = z^{*}_{ic_{sp}} d_i$$

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

To determine an estimate for γ and β of the fuzzy parametric of a sample selection model, one option is to defuzzify the fuzzy observations $\widetilde{W}^{'}_{i_{sp}}$, $\widetilde{X}^{'}_{i_{sp}}$ and $\widetilde{Z}^{*}_{i_{sp}}$. This means converting the 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. The centroid method or the center of gravity method is used to compute the outputs of the crisp value as the center of the area under the curve. Let $W_{ic_{sp}}$, X_{icsp}, and $Z^{*}_{ic_{sp}}$ be the defuzzified values of $\widetilde{W}_{i_{sp}}$, $\widetilde{X}_{i_{sp}}$, and $\widetilde{Z}^{*}_{i_{sp}}$ respectively. The calculation of the centroid method for $W_{ic_{sp}}$, $X_{ic_{sp}}$, and $Z^{*}_{ic_{sp}}$ respectively is via the following formulas:

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

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

and

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

Thus, the crisp values for the fuzzy observation are calculated following the centroid formulas as stated above. To estimate γ_{sp} and β_{sp} with the SPSSM approach, applying the procedure as in Powell, then the parameter is estimated for the fuzzy semi-parametric sample selection model (fuzzy SPSSM). Before obtaining a real value for the fuzzy SPSSM coefficient estimate, first the coefficient estimated values of γ and β are used as a shadow of reflection to the real one. The values of $\hat{\gamma}$ and $\hat{\beta}$ are then applied to the parameters of the parametric model to obtain a real value for the fuzzy SPSSM coefficient estimates of γ_{sp}, β_{sp}, $\sigma_{\varepsilon_{i_{sp}}}$, $u_{i_{sp}}$. The Powell (1987) SPSSM procedure is then executed using the XploRe software.

The Powell SPSSM procedure combines the two-equation structure as shown above but has a weaker assumption about the joint distribution of the error terms:

$$f(\varepsilon_{i_{sp}}, u_{i_{sp}} \mid w_{i_{sp}}) = f(\varepsilon_{i_{sp}}, u_{i_{sp}} \mid w^{'}_{i_{sp}} \gamma).$$

For this reason, it is assumed that the joint densities of $\varepsilon_{i_{sp}}$, $u_{i_{sp}}$ (conditional on $w_{i_{sp}}$) are smooth but unknown functions $f(\cdot)$ that depend on $w_{i_{sp}}$ only through the linear model $w_{i_{sp}}^{'}\gamma$. Based on this assumption, the regression function for the observed outcome z_i takes the following form:

$$\begin{aligned} E(z_i \mid x_{i_{sp}}) &= E(z_{i_{sp}}^{*} \mid w_{i_{sp}}, d_{i_{sp}}^{*} > 0) \\ &= w_{i_{sp}}^{'}\gamma + E(u_{i_{sp}} \mid w_{i_{sp}}, x_{i_{sp}}^{'}\beta > -\varepsilon_{i_{sp}}) \\ &= w_{i_{sp}}^{'}\gamma + \lambda(x_{i_{sp}}^{'}\beta) \end{aligned}$$

where $\lambda(\cdot)$ is an unknown smooth function. The Powell idea of SPSSM is based upon two observations, i and j, with conditions $w_{i_{sp}} \neq w_{j_{sp}}$ but $w_{i_{sp}}^{'}\gamma = w_{j_{sp}}^{'}\gamma$. With this condition, the unknown function $\lambda(\cdot)$ can be differenced out by subtracting the regression functions for i and j:

$$\begin{aligned} E(z_{i_{sp}}^{*} \mid w = w_{i_{sp}}) - E(z_{j_{sp}}^{*} \mid w = w_{j_{sp}}) \\ &= (w_{i_{sp}} - w_{j_{sp}})'\gamma + \lambda(x_{i_{sp}}^{'}\beta) - \lambda(x_{j_{sp}}^{'}\beta) \\ &= (w_{i_{sp}} - w_{j_{sp}})'\gamma \end{aligned}$$

This is the basic idea underlying the γ estimator proposed by Powell (1987). Powell's procedure is from the differences, regress z_i on differences in $w_{i_{sp}}$, as the concept of closeness with two estimated indices $x_{i_{sp}}^{'}\hat{\beta}$ and $x_{j_{sp}}^{'}\hat{\beta}$ (hence $\lambda(x_{i_{sp}}^{'}\hat{\beta}) - \lambda(x_{j_{sp}}^{'}\hat{\beta}) \approx 0$). Thus, γ can be estimated by a weighted least squares estimator:

$$\hat{\gamma}_{Powell} = \left[\binom{n}{2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\hat{\varpi}_{ij}N(w_{i_{ap}} - w_{j_{sp}})(w_{i_{ap}} - w_{j_{ap}})^{'}\right]^{-1} \times$$
$$\left[\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\hat{\varpi}_{ij}N(w_{i_{ap}} - w_{j_{ap}})(z_{i_{ap}} - z_{j_{ap}})\right]$$

Where weights $\hat{\varpi}_{ij}N$ are calculated by $\hat{\varpi}_{ij}N = \frac{1}{h}\kappa\left(\frac{x_{i_{sp}}^{'}\hat{\beta} - x_{j_{sp}}^{'}\hat{\beta}}{h}\right)$ with a symmetric kernel function $\kappa(\cdot)$ and bandwidth h. As shown in earlier equations, this tacitly assumes that $\hat{\beta}$ has previously been obtained as an estimate β. Based on this assumption, a single index model is obtained for the decision equation in place of the probit model (probit step) in the parametric case:

$$P(d_i(d_i^{'} > 0 \mid x) = 1) = g(x_i^{'}\beta)$$

where $g(\cdot)$ is an unknown, smooth function. Using this and given $\hat{\beta}$, the second step consists of estimating γ. Executing the Powell procedure by XploRe software 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 estimate for the first-step index $x_{i_{sp}}^{'}\hat{\beta}$, 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 (not covered in this study) is used for estimating the slope coefficients. For fuzzy PSSM, the above procedure is followed, and then another set of crisp values $W_{ic_{sp}}$, $X_{ic_{sp}}$ and $Z_{ic_{sp}}$ are obtained. Applying the α-cut values on the triangular membership function of the fuzzy observations $\widetilde{W}_{i_{sp}}$, $\widetilde{X}_{i_{sp}}$ and $\widetilde{Z}_{i_{sp}}$ 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 applied. The parameters of the fuzzy SPSSM are estimated. From the various fuzzy data, comparisons will be made on the effect of the fuzzy data and α-cut with original data on the estimation of the SPSSM.

Data Description

The data set used for this study is from the 1994 Malaysian Population and Family

Survey (MPFS-94). This survey was conducted by National Population and Family Development Board of Malaysia under the Ministry of Women, Family and Community Development Malaysia. The survey was specifically for married women, providing relevant and significant information for the problem of married womens' status regarding wages, educational attainment, household composition and other socioeconomic characteristics. The original MPFS-94 sample data comprised 4,444 married women. Based on the sequential criteria (Mroz, 1984) the analyses were limited to the completed information provided by married women; in addition, respondents whose information was incomplete (for example, no recorded family income in 1994, etc.), were removed from the sample.

The resulting sample data consisted of 1,100 married women, this accounted for 39.4% who were employed, the remaining 1,692 (60.6%) were considered as non-participants. The data set used in this study consisted of 2,792 married women. Selection rules (Martins, 2001) were applied to create the sample criteria for selecting participant and non participant married women on the basis of the MPFS-94 data set, which are as follows:

a) Married and aged below 60;
b) Not in school or retired;
c) Husband present in 1994; and
d) Husband reported positive earnings for 1994.

Study Variables

Following the literature (see Gerfin, 1996; Martins, 2001; Christofides, et al., 2003), the model employed in this study consists of two equations or parts. The first equation - the probability that a married women participates in the labor market - is the so-called participation equation. Independent variables involved are: AGE (age in years divided by 10), AGE2 (age squared divided by 100), EDU (years of education), CHILD (the number of children under age 18 living in the family), HW (log of husband's monthly wage). The standard human capital approach was followed for the determination of wages, with the exception of potential experience. The potential experience variable in the data set was calculated by age-edu-6 rather than actual work experience. In order to manage these problems a method advanced by Buchinsky (1998) was used. Instead of considering the term $Q_w = \xi_1 EXP + \xi_2 EXP^2$ in the wage equation i.e., *EXP* is the unobserved actual experience, we use the alternative for women's time is child rearing and the home activities related to child rearing, then the specification of Q_z given by:

$$\begin{aligned} Q_z &= \gamma_1 PEXP + \gamma_2 PEXP2 \\ &\quad + \gamma_3 PEXPCHD + \gamma_4 PEXPCHD2 \end{aligned}$$

The second equation called the wage equation. The dependent variable used for the analysis was the log hourly wages (z). While, the independent variables were EDU, PEXP (potential work experience divided by 10), PEXP2 (potential experience squared divided by 100), PEXPCD (PEXP interacted with the total number of children) and PEXPCHD2 (PEXP2 interacted with the total number of children). Both the participation and wage equations were considered as the specification I and II respectively, that is, the most basic one of SSM.

According to Kao and Chin (2002), the regression parameters (β, γ) should be estimated from the sample data and, if some of the observations in the equation X_{ij} and Y_i are fuzzy, then they fall into the category of fuzzy regression analysis. For the data used in this study, it was assumed that uncertainty was present, therefore, instead of crisp data, fuzzy data are more appropriate. In the participation equation, fuzzy data was used for the independent variables (x): AGE (age in year divided by 10), AGE2 (age square divided by 100) HW (log of husband's monthly wage). For the wage equation, fuzzy data used for the dependent variable was the log hourly wages (z) while the independent variables (x) for fuzzy data involved the variables PEXP (potential work experience divided by 10), PEXP2 (potential experience squared divided by 100), PEXPCD (PEXP interacted with the total number of children) and PEXPCHD2 (PEXP2 interacted with the total number of children). In our study, the observations in the fuzzy participation and fuzzy wage equations involved fuzzy and non-fuzzy data, i.e. a mixture between fuzzy and non-fuzzy data, thus the variables fall into the category of fuzzy data (Kao and Chyu, 2002). For instance, the exogenous variables *AGE, AGE*2 and *HWS* in the participation and the variables *PEXP, PEXP*2, *PEXPCHD* and *PEXPCHD*2 in the wage equations are in the form of fuzzy data. These fuzzy exogenous variables are denoted as $A\widetilde{G}E$, $A\widetilde{G}E2$ $H\widetilde{W}S$ and $P\widetilde{E}XP$, $PE\widetilde{X}P2$, $PEX\widetilde{P}CHD$, $PEX\widetilde{P}CHD2$, respectively. In accord with general sample selection model, the exogenous variables *EDU* and *CHILD* in the participation and the exogenous variable *EDU* in the fuzzy wage equation are considered as non-fuzzy data. However *EDU* and *CHILD* are considered as fuzzy data.

Results

A semi-parametric estimation obtained due to the so-called curse of dimensionality and asymptotic distribution is unknown. Here the results that applied to the most basic estimators are presented; that is, the participant and wage equation of the DWADE estimator and the Powell estimator, respectively. Both estimators are consistent with $\sqrt{n}-$consistency and asymptotic normality.

Participation Equation in the Wage Sector

The participation equation using the DWEDE estimator is presented in Table 1 along with FSPSSM results for comparison purposes. The first column used the DWADE estimator with bandwidth values h = 0.2 without the constant terms. The DWADE estimator shares the ADE estimator of the semi-parametric sample selection model (SPSSM). This is followed by the fuzzy semi-parametric sample selection model (FPSSM) with $\alpha-$cuts 0.0, 0.2, 0.4, 0.6 and 0.8 respectively. At first the estimate coefficient suggests that all variables except AGE are significant (significantly and negatively estimated coefficient on AGE2 and CHILD, while a positive and significant coefficient was estimated for EDU and HW). However, only CHILD shows a statistically significant effect at the 5% level – an unexpected and important result. Although in the conventional parametric model, it appears together with EDU, in the context of SPSSM, only estimates of the CHILD effect appears to be significantly relevant, which is more aligned with economic theory.

For comparison purposes, the FSPSSM was used. The estimated coefficient gives a similar trend with the SPSSM (i.e., significant for variables AGE2, EDU, CHILD and HW). The results show a significant and positive coefficient estimate for EDU and HW, and a significant but negative estimated coefficient on AGE2 and CHILD. In the FSPSSM context, the CHILD coefficient appears to be statistically significant at the 5% level. This is an interesting finding and it should be pointed out that using this approach the standard errors for the parameter were much smaller when compared to those in conventional SPSSM. This provides evidence that this approach is better in estimating coefficients and provides a considerable efficiency gain compared to those in the conventional semi-parametric model. In addition, the coefficient estimated from FSPSSM was considerably close to the coefficient estimated with conventional SPSSM. Hence, the coefficient estimated from FSPSSM is consistent even though it involves uncertain data.

The Wage Equation in the Wage Sector

The wage equation using the Powell estimator of SPSSM is presented in Table 2 with FSPSSM results for comparison purposes. The first column used the Powell estimator with bandwidth values h = 0.2 without the constant terms. The other columns show results given by the fuzzy semi-parametric sample selection model (FPSSM) with $\alpha-$cuts 0.0, 0.2, 0.4, 0.6 and 0.8 respectively.

At first the coefficient estimate suggested that the whole variable was significant (significant and negatively estimated coefficient on EDU, PEXP2 and PEXPCHD, while a positive and significant coefficient was estimated for PEXP and PEXPCHD2). As the estimated coefficient, the results for whole variable statistical significance at the 5% level resulted in a significant result. The results reveal significant differences between the SPSSM compared to the PSSM method of correcting sample selectivity bias. This increased the results obtained in SPSSM where not all variables in PSSM contributed significantly regarding married women involved in wage sectors.

Table1: Semi-Parametric and Fuzzy Semi-Parametric Estimates for the Participation Equation

Participation Equation	Coefficients					
	DWADE	Fuzzy Selection Model				
		α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 0.0
AGE	−0.002048 (1.233)	−0.0015393 (1.150)	−0.0043978 (1.151)	−0.0015934 (1.234)	−0.0016184 (1.232)	−0.001642 (1.232)
AGE2	−0.00016099 (0.1754)	−0.00016584 (0.1624)	−0.00020722 (0.1627)	−0.00016629 (0.1763)	−0.00016651 (0.1765)	−0.00016673 (0.1767)
EDU	0.00034766 (0.02116)	0.00023044 (0.02015)	0.00011323 (0.02015)	0.00023044 (0.02115)	0.00023044 (0.02062)	0.00023044 (0.02062)
CHILD	−0.0039216* (0.06573)	−0.0044301* (0.06341)	−0.0048986* (0.0634)	−0.0044301* (0.06571)	−0.0044301* (0.06485)	−0.0044301* (0.06484)
HW	0.044008 (0.1632)	0.050262 (0.1402)	0.05597 (0.1396)	0.049549 (0.1485)	0.049189 (0.1437)	0.048832 (0.1432)

*5% level of significance

Table 2: Semi-Parametric and Fuzzy Semi-Parametric Estimates for the Wage Equation

Wage Equation	Coefficients					
	Powell	Fuzzy Selection Model				
		α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 0.0
EDU	−0.0112792 (0.005262)	−0.0109003 (0.005258)	−0.010939 (0.005258)	−0.011346 (0.005259)	−0.011385 (0.005259)	−0.0114256 (0.005258)
PEXP	0.544083* (0.1099)	0.540864* (0.1096)	0.538776* (0.1094)	0.534385* (0.1093)	0.532247* (0.1092)	0.530069* (0.109)
PEXP2	−0.160272* (0.02633)	−0.159762* (0.0263)	−0.159524* (0.0263)	−0.158781* (0.02632)	−0.158525* (0.02632)	−0.158259* (0.02632)
PEXPCHD	−0.161205* (0.02453)	−0.159863* (0.02453)	−0.159583* (0.02455)	−0.15889* (0.02459)	−0.158584* (0.02461)	−0.158262* (0.02463)
PEXPCHD2	0.046591* (0.008485)	0.0463242* (0.008485)	.0462221* (0.008493)	0.0458118* (0.008508)	.0457004* (0.008511)	0.0455835* (0.008517)

*5% level of significance

For comparison purposes it was then applied with the FSPSSM. The estimated coefficient was significant for all variables. The results show significant and positive coefficient estimates for PEXP and PEXPCHD2, significant but negative estimated coefficients on EDU, PEXP2 and PEXPCHD. The coefficient for all variables appears to be relevant with statistical significance at the 5% level. It should be noted that, the standard errors for the parameter EDU, PEXP and PEXP2 were much smaller when compared to those in the conventional SPSSM. This provides evidence that this method is considerably more efficient than the conventional semi-parametric model. The coefficient estimated obtained from FSPSSM is also considerably close to the coefficient estimated via conventional SPSSM. In other words, applying FSPSSM, the coefficient estimated is consistent even though the data may contain uncertainties.

Conclusion

For comparison purposes of the participant equation, the estimated coefficient and significant factor gives a similar trend as the SPSSM. However, an interesting finding and the standard errors of the coefficient estimate for the FSPSSM are smaller when compared to the conventional SPSSM. This is evidence that the FSPSSM approach is much better in estimate coefficient and results in considerable efficiency gain than the conventional semi-parametric model. The coefficient estimate obtained was also considerably close to the coefficient estimate of conventional SPSSM, hence providing evidence that the coefficient estimate is consistent even when data involves uncertainties.

The wages equation is similar to the PSSM in terms of the coefficient estimation and significance factors. However, applying the FPSSM resulted in the most significant results when compared to the PSSM, the coefficient estimates of most variables had small standard errors. The rest is considerably close to the standard error of SPSSM. As a whole, the FSPSSM gave a better estimate compared to the SPSSM. In terms of consistency the coefficient estimate for all variables of FSPSSM were not much different to the coefficient estimates of SPSSM even though the values of the $\alpha-\text{cuts}$ increased (from 0.0 to 0.8). In the other words, by observing the coefficient estimate and consistency, fuzzy model (FPSSM) performs much better than the model without fuzzy (PSSM) for the wage equation.

References

Andrews, D. W. K. (1991). Asymptotic normality of series estimation for nonparametric and semiparametric regression models. *Econometrica*, *59*, 307-345.

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.

Chen, T., & Wang, M. J. J. (1999). Forecasting methods using fuzzy concepts. *Fuzzy Sets and Systems*, *105*, 339-352.

Christofides, L. N., Li, Q., Liu, Z., & Min, I. (2003). Recent two-stage sample selection procedure with an application to the gander wage gap. *Journal of Business and Economic Statistics*, *21*(3), 396-405.

Cosslett, S. (1991). Semiparametric estimation of a regression models with sample selectivity. In Barnett, W. A., Powell, J., & Tauchen, G. E. (*Eds.*), 175-198. *Nonparametric and semiparametric estimation methods in econometrics and statistics*. Cambridge, MA: Cambridge University Press.

Gallant, R., & Nychka, D. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*, 363-390.

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

Härdle, W., Klinke, S., & Müller, M. (1999). Xplore learning guide. Berlin: Springer-Verlag.

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*, 475-492.

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

Ichimura, H., & Lee, L. F. (1990). Semiparametirc least square estimation of multiple index models: Single equation estimation. In Barnett, W. A., Powell, J., & Tauchen, G. E. (*Eds.*), 175-198. *Nonparametric and semiparametric estimation methods in econometrics and statistics*. Cambridge, MA: Cambridge University Press.

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

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

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

Lee, M-J., & Vella, F. (2006). A semi-parametric estimator for censored selection models with endogeneity. *Journal of Econometrics*, *130*, 235-252.

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*, 23-39.

Miceli, D. (1998). *Measuring poverty using fuzzy sets*. Discussion paper no.38; National centre for social and economic modeling, University of Canberra.

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.

Newey, W. (1988). *Two step series estimation of sample selection models*. Department of Economics, MIT working paper no. E52 - 262D, 1-17.

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

Powell, J., Stock, J. H., & Stoker, T. M. (1989). Semi-parametric estimation of index coefficients. *Econometrica*, *57*, 1403-1430.

Safiih, L. M. (2007). *Fuzzy semi-parametric of a sample selection model*. Ph.D. dissertation. University Science of Malaysia.

Robinson, P. M. (1988). Root-N consistent semi-parametric regression. *Econometrica*, *56*, 931-954.

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

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, *54*, 1461-1481.

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

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

Yen, K. K., Ghoshray. S., & Roig. G. (1999). A linear regression model using triangular fuzzy number coefficients. *Fuzzy Sets and Systems*, *106*, 167-177.

Zadeh, L. A. (1965). Fuzzy Sets and Systems. In Fox, J. (*Ed.*), *System Theory*. Brooklyn, NY: Polytechnic Press.