American Journal of Applied Sciences 6 (10): 1845-1853, 2009
ISSN 1546-9239

Fuzzy Parametric of Sample Selection Model Using Heckman Two-Step Estimation Models

[1]M.S. Lola, [2]A.A. Kamil, and [1]M.T. Abu Osman
[1]Department of Mathematics, Faculty of Science and Technology,
KUSTEM 21030 Kuala Terengganu, Terengganu, Malaysia
[2]School of Distance Education, University Sains Malaysia,
11800 USM Penang, Malaysia

Abstract: Problem statement: It is well known that, the standard approach to estimating a sample selection models shows an inconsistent estimation results if the distributional assumption are incorrect. **Approach:** An important progress in the last decade to develop an alternative to overcome the deficiency is through the used of semi-parametric method. However, the usage of semi-parametric approach still does not cover the deficiency of the model. **Results:** We introduced a fuzzy membership function for solving uncertainty data of a sample selection model and employed method for sample selection models, that is, the two-step estimators to estimate a model of the so-called the self-selection decision. Fuzzy Parametric of Sample Selection Model (FPSSM) is builds as a hybrid to the conventional parametric sample selection model. **Conclusion/Recommendations:** The result showed that as a whole, the FPSSM give a better estimate and consistent when compared to the Parametric of Sample Selection Model (PSSM). This application demonstrate that the proposed fuzzy modeling approach was quite reasonable and provides an important and significant finding compared with conventional method especially in terms of estimation and consistency.

Key words: Uncertainty, ambiguity, sample selection model, crisp data, membership function

INTRODUCTION

Sample selection is an econometric model that has been found interesting application in empirical studies. Sample selection model, also known as 'self-selection' or 'selectivity' gives a good prior knowledge about relationships and provides an ideal way to incorporate expert judgment and quantitative information. Generally, selection can occur in a linear regression model when data on the dependent variable are missing non-randomly conditional on the independent variables. But, when observations are selected which are not independent of the outcome variables of the study, this sample selection leads to biased inferences. Problems arise when the researcher fails to observe a random sample of a population of interest. With this, model with parametric distributions is subject to distributional misspecifications and tends to result in inconsistent estimates.

Since a random sample does not mirror the true population member's, a lot of discussion have been highlighted especially in the context of labor economics concerning labor force participation, wages and earnings centers on the wage offer distribution, union membership, evaluation of the benefits of social programs[1,5,6,10,11,14]. Their researches discuss the problem of sample selection bias in the context of the decision by women to participate in the labor force or not. The observed distribution represents only one part of the wage offer distribution but being rejected by the other part by the job seekers as unacceptable. Thus, this scenario of estimation procedures may involve certain biases when applied to the secondary labor groups for example married women, teenagers and the aged. Martins[12] discussed a central problem in estimating married women's labor supply functions, in that no market wage is observable for women who do not work. Observation using women, who work to form the sample on which to base the estimation, would cause sample selection bias.

The purpose this study is to introduce 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.

MATERIALS AND METHODS

The Parametric Sample Selection Model (PSSM): Roy's[17] is a good starting point for a formal discussion

Corresponding Author: A.A. Kamil, School of Distance Education, University Sains Malaysia, 11800 USM Penang, Malaysia
Tel: +604 653 4565 Fax: +604 657 6000

on the sample selection problem in the economic literature through "Some Thought on the Distribution of Earnings". He discusses the optimizing choice of 'professions' selecting between fishing and hunting (rabbits) based on their comparative advantage or based on their productivity in each.

The conventional sample selection model (PSSM) as proposed by Heckman[7] can be written as the form:

$$\begin{aligned} &z_i^* = w_i'\gamma + \varepsilon_i \qquad i = 1,...,N \\ &d_i = 1 \text{ if } d_i^* = x_i'\beta + u_i > 0, \\ &d_i = 0 \text{ otherwise} \quad i = 1,...,N \\ &z_i = z_i^* d_i \end{aligned} \tag{1}$$

Where:

d_i and z_i = Dependent variables
x and w = Vectors of exogenous remaining variables
γ and β = Unknown parameter vectors
ε_i and u_i = Zero mean error terms

The standard approach is to assume that (ε_i, u_i) follow a bivariate normal distribution and then applied to the maximum likelihood estimation or a two-stage estimation procedure purposed by Heckman[8]. Firstly, how to estimate γ and β consistently from the data {z_i, d_i, x_i, w_i}, i = 1,…n.? In general, both the error terms are correlated, since that the regression of z on w for the selected sample will not give consistent estimates of γ. It is well known that the consistency of those estimators depends on the assumption of bivariate normality. For a random sample from the population it is observed that d_i, x_i and w_i. If and only if, observation of d_i = 1 then, we observed z_i. This sample selection models in (1) consist of two equations or parts; the first structural part, embodying the desired population relationship or is the equation of primary interest and second, the selection part or is the reduced form takes account of the non-representative nature of the present non-random sample. Following the literature[4,12], the identification purpose, the variable x_i contains at least one variable which does not appear in variable w_i. The structural part describes the relation between an outcome in interest z_i^* and a vector of covariates w_i and the selection equation describing the relation between a binary participation decision d_i^* and another vector of covariates x_i.

In this study, a classical parametric approach was first considered, to estimate the parameters γ and β in the model (1) which specify the joint distribution of the error terms ε and u as bivariate normal and then estimate this parameter along the nuisance parameters of the assumed distribution by maximum likelihood. By allowing the full maximum likelihood estimates in (1) can be computationally cumbersome[15]. To overcome, another most frequently used in practice approached proposed by Heckman through estimating the parameters in a two stages estimator. There are two popular approaches to estimating the sample selection models under this distribution assumption: the widely use procedure of Heckman Maximum Likelihood Estimation[6] and Heckman two step[8].

The more frequently employed method for sample selection models are the two-step estimators introduced by Heckman[8]. In terms of efficiency, it is the second-best alternative to maximum likelihood. The purpose of this model is to estimate a model of the so-called the self-selection decision. In this estimator, the first step, estimating the binary selection equation through probit over the full sample i = 1…N in order to obtain estimates of $\hat{\beta}$. Refers model in (1) and considers the bivariate normal distribution for the error terms, implying independence of the errors and regressors[16]:

$$\begin{aligned} &z_i = w_i'\gamma + \varepsilon_i \\ &d_i = 1(x_i'\beta + u_i > 0) \\ &(\varepsilon_i, u_i) \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & 1 \end{pmatrix}\right) \end{aligned} \tag{2}$$

with z_i only observed for d_i = 1 and since the third row of model in (2) by assumption where σ_u^2 is normalized to 1 as it is not identified in the binary response model, ε_i, u_i are assumed independently and identically distributed and are independent of x_i, then, the model in (2) can be rewritten to:

$$\begin{aligned} &z_i = w_i'\gamma + \sigma_{\varepsilon u}\lambda(x_i'\beta) + \xi_i \\ &d_i = 1(x_i'\beta) + u_i > 0) \end{aligned} \tag{3}$$

where, $\lambda(\cdot)$ is the inverse Mill's ratio, imply by the bivariate normality of (ε_i, u_i):

$$\lambda(\cdot) = \frac{\varphi(\cdot)}{\Phi(\cdot)} \tag{4}$$

$\phi(.)$ and $\Phi(.)$ are the univariate probability density and cumulative distribution function respectively of the standard normal distribution N(0,1) and $\sigma_{\varepsilon u}$ is the covariance between ε and u. The parameters model γ, β

and $\sigma_{\varepsilon u}$ then can be consistently estimated by the following two-step procedure as proposed by Heckman[8].

Probit step: Estimate of β by fitting the probit model $\text{Prob}\{d=1 \mid x\} = E[d \mid x] = \Phi(x'\beta) = F(x'\beta)$ using the full sample 1…N to obtain estimates of $\hat{\beta}$. Then $\hat{\lambda}_i = \frac{\varphi(x_i'\hat{\beta})}{\Phi(x_i'\hat{\beta})}$ can be calculated for each observation with $d_i = 1$ (sub-sample 1…n) and inserted into the structural equation for λ as the additional regressor:

$$z_i = w_i'\gamma + \sigma_{\varepsilon u}\hat{\lambda}_i + \xi_i' \quad (5)$$

OLS step: Using only observations with $d_i = 1$ to estimate the regression function $E(z_i \mid x_i) = x_i'\beta + \sigma_{\varepsilon u} \cdot \frac{\varphi(z_i'\gamma)}{\Phi(z_i'\gamma)}$ by an OLS regression of the observed z_i on x_i and $\frac{\varphi(x_i'\hat{\beta})}{\Phi(x_i'\hat{\beta})}$ where $\hat{\beta}$ is the first step estimate of β.

Winship and Mare[18] indicates that the precision of the estimates in (3) is sensitive to the variance of λ and collinearity between w and λ. The variance of λ is determined by how effectively the probit equation at the first stage predicts which observations are selected into the sample. That means, the better the estimation the greater the variance of λ and the more precise the estimates will be. While, collinearity will be determined in part by the overlap in variables between w and λ.

Fuzzy modeling: Fuzzy modeling used in this study is more on the computational framework which is based on the concepts of fuzzy sets. In the development of PSSM modeling using fuzzy concept, it is considered the basic configuration of fuzzy modeling i.e., fuzzification, fuzzy environment and defuzzification. At the fuzzification stage, an element of real-valued input variables is converted in the universe of discourse into value of membership fuzzy set. In this approach, triangular fuzzy number is used over all observations. The α-cut method with increment value of 0.2 started with 0 up to 0.8 is then applied to the triangular membership function. From the α-cut method, a lower and upper bound for each observations is obtained (x_i, w_i and z_i^*) which is defines as:

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

and

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

Followed by their memberships functions respectively defined have the form as:

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

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

and

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

Based on the condition and problems of the model occurring in this study which involves uncertainties, fuzzy environment such as fuzzy sets and fuzzy number are more appropriates as the processing of the fuzzified input parameters. To find an estimate for γ and β of the fuzzy parametric of sample selection model, one idea is to defuzzify the fuzzy observations $\tilde{W}_i$, $\tilde{X}_i$ and $\tilde{Z}_i$. This 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}, X_{ic} and Z_{ic} be the defuzzified values of $\tilde{W}_i$, $\tilde{X}_i$ and $\tilde{Z}_i$ respectively. The calculation of the centroid method for, W_{ic}, X_{ic} and Z_{ic} respectively via the following formula:

$$W_{ic} = \frac{\int_{-\infty}^{\infty} w\mu_{\tilde{w}i}(w)dw}{\int_{-\infty}^{\infty} \mu_{\tilde{w}i}(w)dw} = \frac{1}{3}(W_{il} + W_{im} + W_{iu})$$

$$X_{ic} = \frac{\int_{-\infty}^{\infty} x\mu_{\tilde{x}i}(x)dx}{\int_{-\infty}^{\infty} \mu_{\tilde{x}i}(x)dx} = \frac{1}{3}(X_{il} + X_{im} + X_{iu})$$

$$Z_{ic} = \frac{\int_{-\infty}^{\infty} z\mu_{\tilde{z}i}(z)dz}{\int_{-\infty}^{\infty} \mu_{\tilde{z}i}(z)dz} = \frac{1}{3}(Z_{il} + Z_{im} + Z_{iu})$$

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. This means, each observation considered has variety 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 large size of observation, the upper bound and lower bound of each observation are quite difficult to be obtained.

Consider the following of the conventional parametric of sample selection model by Heckman[7]:

$$z_i^* = w_i'\gamma + \varepsilon_i \qquad i = 1,...,N$$
$$d_i = 1 \text{ if } d_i^* = x_i'\beta + u_i > 0,$$
$$d_i = 0 \text{ otherwise } i = 1,...,N$$
$$z_i = z_i^* d_i$$

Where:

$$(\varepsilon_i, u_i) \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & 1 \end{pmatrix}\right)$$

From the model, since the variance of u_i is not identifiable, it is consider set to 1. Here it is assumed that two independent variables w, x and the dependent variable z_i^* are involved uncertainty and by applying with fuzzy concept, its can be considered as fuzzy variables. Since it is considered that the two variables involved uncertainty then the error terms (ε_i, u_i) of the models are also considered as fuzzy. This scenario follows Kao and Chin[9] i.e., if some of the observations (x_i and w_i) are fuzzy, then it falls into the category of fuzzy regression analysis.

Based on that the above definition and explanation, fuzzy parametric of sample selection model (FPSSM) is builds as a hybrid to the conventional parametric sample selection model is as follows:

$$\tilde{z}_i^* = \tilde{w}_i'\gamma + \tilde{\varepsilon}_i \qquad i = 1,...,N$$
$$d_i = 1 \text{ if } d_i^* = \tilde{x}_i'\beta + \tilde{u}_i > 0,$$
$$d_i = 0 \text{ otherwise } i = 1,...,N$$
$$z_i = z_{ic}^* d_i$$

The terms $\tilde{w}_i$, $\tilde{x}_i$, $\tilde{z}_i^*$, $\tilde{\varepsilon}_i$ and $\tilde{u}_i$ are fuzzy numbers with the membership functions $\mu_{\tilde{w}_i}$, $\mu_{\tilde{x}_i}$, $\mu_{\tilde{z}_i^*}$, $\mu_{\tilde{\varepsilon}_i}$ and $\mu_{\tilde{u}_i}$ respectively. Since the error terms ε_i and u_i are assumed to follow a bivariate normal distribution for parametric of sample selection model, then for the analysis of the fuzzy parametric of sample selection model, it is also assumed that the crisp values for the error terms follow a bivariate normal distribution i.e.:

$$(\varepsilon_{ic}, u_{ic}) \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & 1 \end{pmatrix}\right)$$

Before obtaining a real value of the Heckman two-step coefficient estimate, first an execution of the coefficient estimate values of γ and β as a shadow of reflection to the real one. The value of $\hat{\gamma}$ and $\hat{\beta}$ above is then applied to the parameters of the parametric model to get a real value for the Heckman coefficient estimate of $\gamma, \beta, \sigma_{\varepsilon_i}, \sigma_{u_i}$. Execution through Xplore software, the Heckman two-step procedure is as follows:

- Step 1 by probit model to estimate γ through fitting the probit model i.e.:

  $$P(d(d_i^* > 0 \mid x) = 1 = E[d \mid x] = \Phi(x'\beta) = F(x'\beta)$$

  Through all over the full sample 1…N with the women participate to the labor force $d(d_i^* > 0 \mid x) = 1$ (for our case, the women participate to the labor force) and $d(d_i^* > 0 \mid x) = 0$ (the women non-participate to the labor force).
  At this step, estimating a binary decision equation (participant equation) takes accounts of the non-representative nature of the sample i.e., 1 for participant and 0 otherwise
- Step 2 by OLS to estimate the regression function by using only observations for $d(d_i^* > 0 \mid x) = 1$ i.e.,

  $$E(z_i^* \mid w_i) = w_i'\gamma + \sigma_{\varepsilon u} \cdot \phi(w_i^*\gamma)/\Phi(w_i^*\gamma)$$

  By an OLS regression of the observed z_i on w_i and $\phi(w_i^*\hat{\gamma})/\Phi(w_i^*\hat{\gamma})$, where $\hat{\gamma}$ is the first step estimate of γ.

In this step, estimates of the parameters of an outcome equation (selection part) on which the significant parties interest is centered

From that program, as input is taken from observations on z_i^*, w, x and $d(d_i^* > 0 \mid x) = 1$ (known as q) and returns estimates of $\gamma, \sigma_{\varepsilon_i u_i}$ and β (placed in heckit.b, heckit.s and heckit.g) respectively. The error terms for the decision and outcome equations should be strongly correlated when applying the above equation with simulated real data. Since the real data generation process satisfies the assumption of the PSSM, then coefficient estimates are quite close to the true coefficients. For fuzzy PSSM, follows the above procedure then another set of crisp values W_{ic}, X_{ic} and Z_{ic} is obtained. Applying the α-cut values on the triangular membership function of the fuzzy observations $\tilde{W}_i$, $\tilde{X}_i$ and $\tilde{Z}_i$ with the original observation, fuzzy data without α-cut and fuzzy data with α-cut to estimate the parameters of the fuzzy parametric of sample selection model. Applying the same procedure above, it is then estimated that the parameters of the fuzzy parametric of sample selection model. From the various fuzzy data, comparisons the effect on the estimation of the parameters of the sample selection model of the fuzzy data and α-cut with original data.

Data description and Variables used:

Data description: The data set used for this study is from the Malaysian population and family survey 1994 (MPFS-1994). This survey was conducted by National Population and Family Development Board of Malaysia under Ministry of Women, Family and Community Development Malaysia. This survey was specifically for married women, providing relevant and significant information for the problem of married women status regarding wages, educational attainment, household composition and other socioeconomic characteristics. The original MPFS-94 sample data comprises 4444 married women. Based on the sequential criteria[15] this analysis was limited to the completed information provided by the married women. For those who gave incomplete information, for example incomplete the survey forms don't have children under 3 years old (YCHILD), no recorded family income in 1994, were removed from the sample. The resulting sample data set consisted only 1100 married women, this accounted for 39.4% who were employed and the rest were considered as non-participants amounting to 1692 (60.6%). The whole data sets used in this study consisted of 2792 married women. The selection rules (Martins, 2001) were applied to create the sample criteria of choosing for participant and non participant married women on the basis of the MPFS-94 data set, which are as follows:

- Married and aged below 60
- Not in school or retired
- Husband present in 1994
- Husband reported positive earning for 1994

Variables used in the study: In this study following the literatures[3,4,12], the model consists of two equations or parts. The first equation which is the probability that a married women participates in the labor market the so-called participation equation, The independent variables involved are AGE (age in year divided by 10), AGE2 (age square divided by 100), EDU (years of education), CHILD (the number of children under 18 living in the family), HW (log of monthly husband's wage). The standard human capital approaches was followed for the determination of wages except the potential experience. For the potential experience available in the data set, the calculation was given by age-edu-6 rather then actual work experience. In order to deal with these problems the solution was adopted using a method advanced by Buchinsky[2]. Instead of considering the term $Q_w = \xi_1 EXP + \xi_2 EXP^2$ in the wage equation (actual EXP is the unobserved actual experience), it is assumed that the best alternative use for a woman's time is child rearing (and the home activities related to this task), the specification was included with Q_z given by:

$$Q_z = \gamma_1 PEXP + \gamma_2 PEXP2 + \gamma_3 PEXPCHD + \gamma_4 PEXPCHD2 \quad (6)$$

The second equation called 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 equation were considered as the specification I and II respectively i.e., the most basic one of SSM.

According to Kao and Chin[9], 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 it falls into the category of fuzzy regression analysis. For the data used in this study, it was assumed that the data contained uncertainty, instead of crisp data, fuzzy data are more appropriate. In the participation equation, a fuzzy data was used for the independent variables (x) involve AGE (age in year divided by 10), AGE2 (age square divided by 100) HW (log of monthly husband's wage). For the wage equation, a fuzzy data used for dependent variable was the log hourly wages (z) while the independent variables (x) for fuzzy data involve 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 this study and related to the study of Kao and Chin[9], the data used but did not involved fuzzy so-called a non-fuzzy data. For non-fuzzy data, the variables involved were EDU and CHILD. Since the data are fixed (in terms of integer value) and could not be fuzzified, it was considered fuzzy data as well.

Endogenous variables: In this study "participation equation" was the first dependent variable. This variable is a dichotomous indicator that takes the value 1 if the women participate and 0 otherwise. The category of non-participant in the labor market included individuals who are either self-employed (family business or farming) or exclusively engaged in non-market home production. The highest number of married women participants and non-participants in the labor market were Malay 616 (22.1%) and 1735 (62.1%), Chinese 353 (12.6%) and 717 (25.7%), Indian, 107 (3.8%) and 242 (8.7%) and other races was 24 (0.9%) and 98 (3.6%) respectively.

The second dependent variable was "the log of Hourly Wages (HW)" in the wage equation. In Malaysia remuneration, other then basic wages as an important part of total earning[13]. From the 1994 survey, the Chinese women gave a significantly higher income wages (≥ RM3,000.00 or 1.1%) while equal income wages (0.9%) for Malay and Indian when compared to the wages sector in labor market. The lower hourly wages (≤ RM999) were similar for Malay, Chinese and Indian (96.1, 94.1 and 96.3%) respectively.

Exogenous variables: In this part, the variables for instant AGE, Education (EDU) involved are the participation and the potential experiences of the wage equation are the variables involved in the first equation. The purpose of using the AGE and EDU are to measure general human capital and are expected to have negative effect on the probability of being employed.

Age: The 1994 survey shows that women wage workers (in average) are 18 years old and women non-wage workers are 29 years old. This indicates that the women participating in the labor market are younger then for non-participating women. This result is consistent with the increased important of the wage sector in Malaysia, with reason that individually, the younger women participant in labor market are well educated. The age variable is used to measure general human capital and is expected to have negative effect on the probability of being employed.

The potential experience: This is calculated by age_i-$schooling_i$-6 with women participants (15.4 years) is less when compare to women non-participants (20.8 years). This implies that the women participants in the labor market are influenced by childbearing and child-raising activities. According to the data given and sequences with the total number of children (under 7 years old) of women non-participant are 965 children when compared to the total number of children of women participant is 441 children. Even though, in 1988, the total fertility rate in Malaysia decreased to 3.7% when compared to 6.3% in 1965.

Education: To standardize the measurement of the education attainment was done by the continuous variable i.e., "years of schooling". For information, no indicator of measure was available applied regarding the actual years it took each individual to reach the level completed. For instance, the individual having obtained a post-secondary diploma, the years required were inferred from the degree obtained. From the data reported, only the pre-tertiary grade was completed.

RESULTS

Empirical results: parametric and fuzzy parametric model: The empirical results of the basic specification one are presented for the Heckman two-step approach. These approaches consider the probit estimates for the participation equation as a first step and OLS estimates for the wage equation as the second step. We discuss both the participation and wage equation on the estimated coefficient, the significant effect, consistency and the HH test for PSSM, as well as FPSSM for comparison purposes.

Table 1: Parametric and fuzzy parametric estimates for the participation equation

	Coefficients					
		Fuzzy selection model				
Participation equation	Heckman	$\alpha = 0.8$	$\alpha = 0.6$	$\alpha = 0.4$	$\alpha = 0.2$	$\alpha = 0.0$
Constant	4.01403 (2.939)	4.46099 (2.95)	4.54494 (2.986)	4.62869 (3.021)	5.60029 (3.241)	5.5613 (3.238)
AGE	-0.0077529 (1.603)	-0.0075185 (1.614)	-0.0075319 (1.635)	-0.0075733 (1.656)	-0.0074408 (1.783)	-0.0074725 (1.784)
AGE2	0.37939 (0.2132)	0.37612 (0.2152)	0.37485 (0.2183)	0.37362 (0.2213)	0.37212 (0.2397)	0.3708 (0.2402)
EDU	-0.11004 (0.02288)	-0.10945 (0.02265)	-0.10939 (0.02265)	-0.10929 (0.02256)	-0.10927 (0.02282)	-1.10917 (0.02282)
CHILD	-0.14737* (0.06241)	-0.14563* (0.06146)	-0.14562* (0.06145)	-0.14557* (0.06144)	-0.14422* (0.06243)	-0.14417* (0.06241)
HW	0.040431* (0.1231)	0.039965* (0.1098)	0.039963* (0.1092)	0.039947* (0.1087)	0.039708* (0.113)	0.039689* (0.1125)

*: 5% level of significant

The participation equation in the wage sector: In Table 1 we present the empirical results of the basic specification one for the first step of Heckman two-step approach. The results of this approach consider the probit estimates, then as comparison to the fuzzy parametric of sample selection model.

The first column presents the Parametric Selection Model of Heckman two-step estimates (PSSM). These give generally the probit results on the estimates for the participation in the wage sector. The following column represents a Fuzzy Parametric of Sample Selection Model (FPSSM) with α-cuts 0.0, 0.2, 0.4, 0.6 and 0.8, respectively. For the case of PSSM, the estimation coefficient purpose suggests that the Husband Wage's (HW) shows a significant and positive coefficient estimate. Significantly but negative coefficient estimated on EDU and CHILD (the number of children in the family). To test the presence of selectivity bias into the model, between the errors on the participant equations is through the null hypothesis with no correlation ($\rho = 0$). The results shown that, the family size (measured by the number of children in the family) and HW failed to be rejecting at 5% level of significant. In other words, both are the important significant factors for a women's decision to participant in the labor market. Having a CHILD for married women will be effect the decision to participant or not into labor market. For married women with a small family (number of children $\leq$3) the tendency not to participant into the labor market is high (or 1610 or 57.6% of married women), when compared to the participation of married women with a small family (1058 or 37.9%). While, mean and standard deviation for the married women with high husband income ($\geq$RM3,000.00) the decision of married women to participate in the labor market when less as compare to the married women with low and middle husband income ($\leq$ RM900.00 and RM 1,000.00-RM2,900.00) i.e., 2.586 (0.580) and 2.710 (0.601) respectively.

Applying the FPSSM as a comparison, the results in terms of the coefficient estimation and significant factor show a similar trend with PSSM i.e., the Husband Wage's (HW) show a significant and positive coefficient estimate. But negative coefficient estimated and significantly on EDU and CHILD (the number of children in the family). While, also failed to reject the of CHILD and HW variables with 5% level of significant. Means, both variables still become significant factors for women's decision to participate in the labor market. But, the most significant result by applying the FPSSM is that, the coefficient estimated for variables EDU, CHILD and HW gives a better estimate when compared to the PSSM in terms of the standard error of the coefficient estimate. In terms of consistency, by applying the FPSSM, all variables are consistent even though the α-cuts values increases (from 0.0-0.8), the coefficient estimate are still close to the coefficient estimate of PSSM. In the other words, in terms of coefficient estimate and consistency, fuzzy model (FPSSM) is much better then the model without fuzzy (PSSM) for participant equation.

The wage equation in the wage sector: Table 2 presents the empirical results of the OLS estimates for the wage equation for the Heckman two-step approach. The first column presents the parametric selection model of Heckman two-step estimates (PSSM). These give the probit results on the estimates for the wage regressions. The following columns present a Fuzzy Parametric of the Sample Selection Model (FPSSM) with α-cuts 0.0, 0.2, 0.4, 0.6 and 0.8, respectively. Table 2 shows generally the result for the wage equation. These are surprising, because all variables show a significant, positive coefficient estimate for EDU and PECPCHD2 and negative coefficient estimate for PEXP, PEXP2 and PEXPCHD effect to the women wages. To test the presence of selectivity bias into the model i.e., between the errors on the wage equations is through the null hypothesis with no correlation ($\rho = 0$). As a result, the test failed to reject the PEXP, PEXP2, PEXPCHD and PEXPCHD2 variables with 5% level of significance. In other words, the PEXP, PEXP2, PEXPCHD and PEXPCHD2 variables give a significant effect for the women wages.

Table 4.2: Parametric and Fuzzy parametric estimates for the wage equation

	Coefficients					
		Fuzzy selection model				
Wage equation	Heckman	$\alpha = 0.8$	$\alpha = 0.6$	$\alpha = 0.4$	$\alpha = 0.2$	$\alpha = 0.0$
Constant	-0.122217 (0.1197)	-0.118844 (0.1195)	-0.121264 (0.119)	-0.121302 (0.1187)	-0.118386 (0.1184)	-0.118413 (0.1181)
EDU	1.3971 (0.005285)	1.4609 (0.005274)	1.4609 (0.005274)	1.4695 (0.005281)	1.4551 (0.005282)	1.4599 (0.005282)
PEXP	-0.2033* (0.1102)	-0.20924* (0.11)	-0.20924* (0.11)	-0.21115* (0.1097)	-0.20333* (0.1095)	-0.20433* (0.1093)
PEXP2	-0.03411* (0.02642)	-0.03433* (0.02637)	-0.03433* (0.02637)	-0.034335* (0.02641)	-0.030416* (0.02642)	-0.030446* (0.02642)
PEXPCHD	-0.19861* (0.02452)	-0.20736* (0.02447)	-0.20722* (0.02455)	-0.20708* (0.02457)	-0.21048* (0.02459)	-0.21043* (0.02461)
PEXPCHD2	0.3052* (0.008497)	0.3055* (0.008479)	0.30435* (0.008506)	0.3032* (0.008513)	0.27319* (0.008521)	0.27182* (0.008527)

*: 5% level of significant

For comparison purposes, the FPSSM was applied and the result show a similar results with PSSM of the coefficient estimation and significant factor i.e., significant for all variables with positive coefficient estimate for EDU and PECPCHD2 and negative coefficient estimate for PEXP, PEXP2 and PEXPCHD effect on the women wages. Applying the FPSSM gave the most significant result when compared to the PSSM, the coefficient estimated for variables EDU, PEXP, PEXP2 and PEXPCHD gave a small standard error of the coefficient estimate. For PEXPCHD2 also gave a small standard error but only for 0.8 α-cuts values. As a whole, the FPSSM give a better estimate when compared to the PSSM. The study also looked at the consistency when applying the FPSSM. It was found that the coefficient estimate of FPSSM was not much different to the coefficient estimate of PSSM for all variables even though the values of the α-cuts increased (from 0.0-0.8). In other words, by looking at the coefficient estimate and consistency, fuzzy model (FPSSM) is much betters then the model without fuzzy (PSSM) for wage equation.

DISCUSSION

One of the most significant ideas of this study is quite simple. Previously, almost all the literature in the parametric selectivity model was centered on the concept of an inconsistent estimation results if the distributional assumption were incorrect. Hence, semi-parametric model or nonparametric approach in different perspectives applied to overcome that problem. However, none of them put effort to analyze from the perspective of a fuzzy environment which is more realistic especially when dealing with historical data that contain uncertainty. This study is a platform to enter a new dimension on the fuzzy modeling of SSM (Sample Selection Model). However, further research can consider the new development from fuzzy perspective and paradigm.

CONCLUSION

For comparison purposes, firstly, look at the participant equation. The results show a similar trend with PSSM in terms of the coefficient estimation and significance factor. However, the most significant result appears by applying the FPSSM i.e., the FPSSM a better estimate when compared to the PSSM in terms of the standard error of the coefficient estimate. In terms of consistency, by applying the FPSSM, all variables are consistent even though the α-cuts values increases (from 0.0-0.8), the coefficient estimate are still close to the coefficient estimate of PSSM. In the other words, in terms of coefficient estimate and consistency, fuzzy model (FPSSM) is much betters then the model without fuzzy (PSSM) for participant equation. Secondly; wages equation, also does not have much difference with the PSSM, in terms of the coefficient estimation and significant factor. However, applying the FPSSM gave the most significant result when compared to the PSSM, the coefficient estimated of most the variables gave a small standard error. Only a few show a small standard error against PSSM. As a whole, the FPSSM give a better estimate when compared to the PSSM. The study also looks at the consistency when applying the FPSSM. It was found that the coefficient estimate of FPSSM was not much different to the coefficient estimate of PSSM for all variables even though the values of the α-cuts increased (from 0.0 to 0.8). In the other words, by looking at the coefficient estimate and consistency, fuzzy model (FPSSM) was much betters then the model without fuzzy (PSSM) for wage equation.

ACKNOWLEDGEMENT

The research present here is supported by Fundamental Research Grant Scheme (FRGS) No. 203/PJJAUH/671128, Ministry of Higher Education Malaysia.

REFERENCES

1. Amemiya, T., 1984. Tobit models: A survey. J. Econ., 24: 3-61.
2. Buchinsky, M., 1998. The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. J. Applied Econ., 13: 1-30. DOI: 10.1002/(SICI)1099-1255(199801/02)13:1<1::AID-JAE474>3.0.CO;2-A
3. Christofides, L.N., Q. Li, Z. Liu and I. Min, 2003. Recent two-stage sample selection procedure with an application to the gander wage gap. J. Bus. Econ. Stat., 21: 396-405.
4. Gerfin, M., 1996. Parametric and semiparametric estimation of the binary response model of labor market participation. J. Applied Econ., 11: 321-339.
5. Gronau, R., 1974. Wage comparisons: A selectivity bias. J. Political Econ., 82: 1119-1143.
6. Heckman, J.J., 1974. Shadow price, market wages and labor supply. Econometrics, 42: 679-694.
7. 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. Ann. Econ. Soc. Measur., 5: 475-492. http://ideas.repec.org/h/nbr/nberch/10491.html
8. Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica, 47: 153-161.
9. Kao, C. and C.L. Chin, 2002. A fuzzy linear regression model with better explanatory power. Fuzzy Sets Syst., 126: 401-409.
10. Lewis, H.G., 1974. Comments on selectivity biases in wage comparisons. J. Political Econ., 82: 1145-1155. http://ideas.repec.org/a/ucp/jpolec/v82y1974i6p1145-55.html
11. Maddala, G.S., 1983. Limited-Dependent and Qualitative in Econometrics. Series: Econometric Society Monographs (No. 3), Cambridge University Press, Cambridge, pp: 257-289.
12. Martin, M.F.O., 2001. Parametric and semiparametric estimation of sample selection models: An empirical application to the female labor force in Portugal. J. Applied Econ., 16: 23-39. http://www.jstor.org/stable/2678535
13. Mazumdar, D., 1981. The Urban Labor Market and Income Distribution: A Study of Malaysian. Oxford University Press, New York, 1-375.
14. Neumark, D., 1988. Employers' discriminatory behavior and the estimation of wage discrimination. J. Hum. Resour., 23: 279-295. http://www.jstor.org/stable/145830
15. Newey, W., J.L. Powell and J.R. Walker, 1990. Semi-parametric estimation of selection models: Some empirical results. Am. Econ. Rev., 2: 324-328. http://ideas.repec.org/p/att/wimass/9001.html
16. Powell, J.L., 1994. Estimation of Semi-parametric Models. Handbook of Econometrics, Vol. 4. Elsevier Science Publishers, Amsterdam, pp: 2443-2521.
17. Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxf. Econ. Papers, 3: 135-146. http://oep.oxfordjournals.org/cgi/pdf_extract/3/2/135
18. Winship, C. and R.D. Mare, 1992. Models for sample selection bias. Ann. Rev. Sociol., 18: 327-350. http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.so.18.080192.001551