FUZZY PARAMETRIC SAMPLE SELECTION MODELS OF MARRIED WOMEN FOR NON-PARTICIPATION BY MLE : CASE STUDY THE MPFS-1994

L. Muhamad Safiih[1], Yaya Sudarya Triana[2]
[1]Information Systems Department Faculty of Information Technology,
University of Tarumanagara, Jakarta, Indonesia
[2]Mathematics Department Faculty of Science and Technology,
University Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, Malaysia

[1]safiihmd@umt.edu.my, [2]yst_2000@yahoo.com

Abstract

Models with sample-selection biases are widely used in various fields of economics such as labour economics (see Maddala, Amemiya, and Mroz). The models are usually estimated by Heckman's two-step estimator. However, Heckman's two-step estimator often performs poorly (see Wales and Woodland, Nelson, Paarsch, and Nawata). The data used in this study originated from the survey was conducted by the National Population and Family Development Board of Malaysia under the Ministry of Women, Family and Community Development of Malaysia, called the Malaysian Population and Family Survey 1994 (MPFS, 1994). The survey was conducted through a questionnaire, were randomly and specifically for married women. The data set focus on married women which provides information on wages, educational attainment, household composition and other socioeconomic characteristic. The Original sample data based on Mroz (1987), there are 4444 records married women. It is necessary to use the maximum likelihood method to estimate the models in such cases. For solving uncertainty data of a parametric sample selection model, in this paper needs to consider the models estimation using fuzzy modeling approach, called Fuzzy Parametric Sample Selection Model (FPSSM). Fuzzy Parametric sample selection model (FPSSM) is builds as a hybrid to the conventional parametric sample selection model. Finally, the result showed, FPSSM by Maximum Likelihood Estimator (MLE) estimates of the mean, Standard Deviation (SD).

Keywords: Econometrics, Fuzzy Number, Heckman Two-Step Estimator, Married Women, MLE, Non-participation, Sample Selection Model.

1. Introduction

The majority of the population of Malaysia is the Malays, but their economic power is not equivalent to those positions. In 1970, only holds 1.9% Bumiputra Malaysia's economy, while non-Malays (mainly Chinese) holds 37.4%, and the rest in the hands of foreigners. In 1971 a new economic policy is applied. New economic policy aimed to eliminate poverty and eliminate "racial identification on the economic function" through a "thriving economy"; New Economic Policy target of 30% Bumiputra holding the economy in the next 20 years.

The model of female labour participation and wage equations in labour economics were by Gronau (1974) and Heckman (1976, 1979). The study analysis for the Malaysian Population and Family Survey 1994 (MPFS-1994), the original sample data based on Mroz (1987). Data collected from 4444 women heads of households through questioner. The data used in this study originated from the survey was conducted by the National

Population and Family Development Board of Malaysia under the Ministry of Women, Family and Community Development of Malaysia.

In this study, method of calculating Heckman's two-step estimator is compared using modify of Heckman's two-step estimator (1979) for the α-cuts of triangular fuzzy number. However, Heckman's two-step estimator often performs poorly (see Wales and Woodland, Nelson, Paarsch, and Nawata).

It is necessary to use the maximum likelihood method to estimate the models in such cases. For solving uncertainty data of a parametric sample selection model, in this paper needs to consider the models estimation using fuzzy modeling approach, called Fuzzy Parametric Sample Selection Model (FPSSM).

2. Parametric Sample Selection Model

The discussion of sample selection was started by Roy (1951) in the economic literature, Roy (1951) discussed that the traditional econometric approach to the selection model adopts a more conservative approach and allows for selection on unobservables. Several methods have been proposed to cope with this problem. Later this method extended by Gronau (1974), Heckman (1974) , and Lewis (1974). Issues are discussed regarding the sample selection bias in the context of the decision by women to Participate in the labour force or not (Schafgans, 2000).

The most widely used procedure has been suggested by Heckman (1976, 1979) is a process that describes the employment outcome is implemented and information from this is used, in the second stage, to obtain consistent estimates of the relevant parameters. Heckman (1976) has proposed the PSSM as follows:

$$\begin{aligned} Y_i^* &= \beta_0 + X_i'\beta_1 + u_i \\ d_i &= \begin{cases} 1 & \text{if} \quad d_i^* = W_i'\alpha + v_i > 0 \\ 0 & \text{otherwise} \end{cases} \\ Y_i &= Y_i^* d_i \end{aligned} \tag{1}$$

where

Y_i^* , d_i^*	=	dependent variables
X_i ,W_i	=	vector of exogenous variables
β_0 , β_1 ,α	=	unknown parameter vectors
u_i, v_i	=	error terms

In Equation (1) above are error terms (u, v) which are usually correlated, so that will cause the value estimation of β_0 and β_1 unsatisfactory or inconsistent as a result of the regression equation of the dependent variable Y and independent variable X. To reduce this problem, the approach of the error terms are assumed to follow a normal bivariate distribution. In addition Heckman (1979) proposed the concept of MLE is applied to the SSM to improve the previous result.

According to Schafgans (1996), Markus (1998) and Martins (2001), there are two parts in equation (1). The first part is participation equation (a binary decision equation). The second part is the wage equation (outcome equation or selection part). Independent variable X_i usually contain at least one variable which does not appear in variable W_i. In

the outcome equation describes the relationship between the dependent variable Y_i and independent variable X_i, whereas in the selection equation describes the relationship between the dependent variable di and the independent variable W_i.

Equation (1) used by Newey et al. (1990) for maximum likelihood method. As an alternative of estimation that has been proposed by the Heckman Two-step estimator. For the Sample Selection Model distribution under this assumption the Maximum Likelihood Estimator (MLE) by Heckman's (1974) and Heckman's two step estimator (1979). Heckman's (1979) used by Nawata (1994) using the MLE.

3. Maximum Likelihood Estimation

Heckman (1974) who proposed a MLE is a solution of the first to sample selection bias (Vella, 1998). It required the distributional assumption of the disturbances and Heckman made as follows:

The error terms of u_i and v_i are independently and identically distributed (iid) N(0, Σ)

$$\Sigma = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}$$

and (u_i, v_i) are independent of W_i.

According to (Vella, 1998) that assumption above, a straightforward estimation for the parameter in equation (2.1) is done by maximizing the following average log likelihood function.

$$L = \frac{1}{N}\sum_{i=1}^{N}\left\{ d_i * \ln\left[\int_{-(w_i'\alpha)}^{\infty} \phi_{uv}(y_i - x_i'\beta, v)dv \right] + (1-d_i) * \left[\ln \int_{-(w_i'\alpha)}^{\infty}\int_{-\infty}^{\infty} \phi_{uv}(u,v)dudv \right] \right\} \qquad (2)$$

where ϕ_{uv} is the probability density function for the bivariate normal distribution. This estimation in equation (2) would be simplified if $\phi_{uv} = 0$, would then reduce to the product of two marginal likelihoods. In this study, a product of the likelihood function examining whether d_i was equal to 1 or 0, over the entire N samples. The likelihood function explaining the variation in z_i for the n sub-sample satisfying d_i=1. Alternatively, when $\sigma_{uv} \neq 0$, it is necessary to evaluate double integrals. Moreover, because there is no selection bias when $\sigma_{uv} = 0$ a test of this hypothesis is a test of whether the correlation coefficient ρ_{uv} is equal zero as this parameter captures the dependence between u and v. (Vella, 1998)

4. Fuzzy Modeling

In this research discusses the modified Parametric Sample Selection Models (PSSM) for non-participant by applying a fuzzy concept. The fuzzy modeling is based on the concept of fuzzy sets. There are three phases to discuss of FPSSM, namely fuzzification of parameters, the fuzzy environment and defuzzification. In the fuzzification stage, the value of the input variables (crisp values) is converted into fuzzy values input from the membership of fuzzy sets. The method used in this study is an alpha-cut which

applied to the triangular fuzzy number for the all observation. The alpha-cut method starts from 0.2 to 0.8 with increase of 0.2. Then applied into the triangular membership function.

By using alpha-cut, the result will be obtained from (x, w, y) as follows:

$$\tilde{x}_i = (x_{il}, x_{im}, x_{iu}),\ \tilde{y}_i^* = (y_{il}, y_{im}, y_{iu}) \text{ and } \tilde{w}_i = (w_{il}, w_{im}, w_{iu}) \quad (3)$$

The functions of membership are as below :

$$\mu_{\tilde{x}_i}(x) = \begin{cases} \dfrac{(x - x_{il})}{(x_{im} - x_{il})} & \text{if } x \in [x_{il}, x_{im}] \\ 1 & \text{if } x = x_{im} \\ \dfrac{(x_{iu} - x)}{(x_{iu} - x_{im})} & \text{if } x \in [x_{im}, x_{iu}] \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The same formula for $\mu_{y_i^*}(y)$ and $\mu_{\tilde{w}_i}(w)$ can be written like a formula for $\mu_{\tilde{x}}(x)$

To create the PSSM, then the first step is to change the membership (converting real-triangular fuzzy membership values into a crisp value). A centroid method or the centre of gravity method is used to calculate the outputs of the crisp value as the centre of area under the curve. The values of X_{ic}, Y_{ic}, and W_{ic} be the defuzzified values of $\tilde{X}_i$, $\tilde{Y}_i$ and $\tilde{W}_i$ respectively. The calculation of the centroid method for X_{ic}, Y_{ic}, and W_{ic} formulating are:

$$X_{ic} = \frac{\int_{-\infty}^{\infty} x\mu_{\tilde{x}i}(x)dx}{\int_{-\infty}^{\infty} \mu_{\tilde{x}i}(x)dx} = \frac{1}{3}(X_{il} + X_{im} + X_{iu}) \quad (5)$$

5. Data Description and Used Variables

5.1 Data Description

The data used in this study originated from the survey was conducted by the National Population and Family Development Board of Malaysia under the Ministry of Women, Family and Community Development of Malaysia, called the Malaysian Population and Family Survey 1994 (MPFS, 1994). The survey was conducted through a questionnaire, were randomly and specifically for married women. The data set focus on married women which provides information on wages, educational attainment, household composition and other socioeconomic characteristic.

The Original sample data based on Mroz (1987), there are 4444 records married women. Data used for this study using the Malaysian Population and Family Survey 1994 (MPFS, 1994). Then the selection data which considered, there is only married women with completed information. Uncompleted information is grouped into a file invalid data and complete data is grouped into valid data. Furthermore, the processed data is only valid data. After selection sample data, there are 1850 records of married women to be grouped into valid data. Then the valid data selected again to be grouped into participant and non-

participant data, selection of data based on Martin (2001). The criteria of a sample choosing for participant and non-participant married women (MPFS-94), which are:

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

After selection, there are 62.65% or 1159 persons are classified as participating of married women and 37.35% or 691 persons are classified as non-participating of married women.

5.2 Fuzzy Used Variable

From the data set of the MPFS-1994 which contains a non-participating data, there are some exogenous variables using fuzzy concepts, rules Martin (2001) is used as follows:

$\tilde{AGE}$ (in years) is married women's age divided by 10

$\tilde{AGE2}$ is age squared divided by 100

EDU is educational levels, measured in years of schooling

CHILD is the number of children younger than 18 living in the family

$\tilde{HW}$ is the log of the husband's monthly wage (measure in Ringgit Malaysia)

$\tilde{PEXP}$ is potential work experience, defined as age minus 6 minus years of schooling

$\tilde{PEXP2}$ is potential work experience squared

$\tilde{PEXPCHD}$ is potential work experience times the number of children

$\tilde{PEXPCHD2}$ is potential work experience squared, times the number of children

$\tilde{HWW}$ or $\tilde{Y}$ is the log women's hourly wage rate (measured in Ringgit Malaysia)

d is the indicator of labor market nonparticipation

5.3 Endogenous Variables

The category of non-participant in the labour market included individuals either self-employed (family business or farming) or exclusively engaged in non-market home production (Schafgans, 1996). The highest number of married women participants and non-participants in the labour market were Malay: 616 (22.1%) and 1735 (62.1%), respectively, Chinese: 353 (12.6%) and 717 (25.7%), respectively, Indian: 107 (3.8%) and 242 (8.7%) respectively and the other races were 24 (0.9%) and 98 (3.6%) respectively.

The first dependent variable, "non-participant", is a dichotomous indicator that equal 1 if an individual is a non-participant and 0 if not. The second dependent variable is "the log of hourly wages-(HW)" in the wage equation. In Malaysia remuneration, other than basic wages, for instance allowance, bonus, etc, are an important part of total earning (Mazumdar, 1991). Schafgans (1996), therefore, mention that bonuses and payments in-kind (for instance food, housing, etc) are included in the computational of hourly wages.

5.4 Fuzzy Exogenous Variables

Exogenous variables are the variables that enter into the second group, which contained both in the nonparticipation and outcome equations. For example, the variable EDU appear in the nonparticipation and outcome equations, whereas, the variables AGE and potential experiences only appears in the equal nonparticipation and outcome equations, respectively. Details of the exogenous variables are as follows:

$\tilde{AGE}$

The average age of married women non-participants are 33.72 years old, while the age of married women participants are 33.24 years old. This result is in accordance with

Schafgans (1996) in which women participants on average younger than women non-participant women. These results indicated that it is consistent with the importance of increasing wage sector in Malaysia, particularly among the younger educated individuals.

EDU
EDU is the educational levels. This variable measures the school year are required to obtain the highest grade completed. Its measured by continuous variables. There is no measure available regarding the actual year is needed for each individual to achieve the level completed (Schafgans, 1996). The average edu of married women non-participants are 7.35, while the edu of married women participants are 10.96.

$\widetilde{\text{PEXP}}$
The average PEXP for non-participants of married women are 23,85 and while the PEXP of married women participants are 15.43 years.

6. Fuzzy Participation Equation and Fuzzy Wage Equation

According to Greene (1997) model consist of two equation. The first equation is called the participation equation. i.e. the probability of married women non-participating in the labour market. The independent variables consist of $\widetilde{\text{AGE}}$ (age in years divided by 10), $\widetilde{\text{AGE2}}$ (age squared divided by 100), EDU (years of education), CHILD(the number of children under 18 living in the family), and $\widetilde{\text{HW}}$ (log of monthly husband's wage). The dependent variable in the participant equation is a dummy variable that takes the value 1 if the woman non-participate and zero otherwise. Wages is determined by standard human capital approach, The potential experience (given by age-edu-6) available in dataset. Buchinsky (1998) is solution to deal with this problems. The wage equation which is:

$$z_i = \beta_0 + \beta_1 \widetilde{\text{AGE}} + \beta_2 \widetilde{\text{AGE2}} + \beta_3 \text{EDU} + \beta_4 \text{CHILD} + \beta_5 \widetilde{\text{HW}} + \varepsilon \qquad (6)$$

where

$d_i = 1\ (z_i \leq 0)$

The second equation is called the wage equation. The explanatory variables used in the wage equation are as follows: EDU, $\widetilde{\text{PEXP}}$, $\widetilde{\text{PEXP2}}$, $\widetilde{\text{PEXPCHD}}$ and $\widetilde{\text{PEXPCHD2}}$
The fuzzy wage equation as below:

$$z_j = \beta_0 + \beta_1 \widetilde{\text{PEXP}} + \beta_2 \widetilde{\text{PEXP2}} + \beta_3 \widetilde{\text{PEXPCHD}} + \beta_4 \widetilde{\text{PEXPCHD2}} \qquad (7)$$

where

$d_j = 1\ (z_j \leq 0)$

The independent variables are EDU, $\widetilde{\text{PEXP}}$ (potential experience divided by 10), $\widetilde{\text{PEXP2}}$ (potential experience squared divided by 100), $\widetilde{\text{PEXPCHD}}$ ($\widetilde{\text{PEXP}}$ interacted with the total number of children) and $\widetilde{\text{PEXPCHD2}}$ ($\widetilde{\text{PEXP2}}$ interacted with the total number of children). The dependent variable used for the analysis was the long hourly wages (z).

If Participate = 1, the outcome equation is observed, as follow:

$$\text{Ln}(Y) = \beta_0 + \beta_1 \text{PE}\tilde{X}\text{P} + \beta_2 \text{PE}\tilde{X}\text{P2} + \beta_3 \text{PEX}\tilde{P}\text{CHD} + \beta_4 \text{PEX}\tilde{P}\text{CHD2} + \varepsilon \qquad (8)$$

In this study are EDU and CHILD. Assumed data used in this study contained uncertainty, instead of crisp data, therefore data are more appropriate. In the participation equation, fuzzy data used for the independent variables (x). For the outcome equation the fuzzy data that was used for dependent variables was the log hourly wage (z).

7. Result

Table 1. Fuzzy Participation Equation

Model	Variables	Fuzzy Participation Equation							
		α = 0.2		α = 0.4		α = 0.6		α = 0.8	
		Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.
Two-Step	age	0.4773	0.2068	0.4184	0.1791	0.3534	0.1558	0.3058	0.1466
	age2	-0.0865	0.0349	-0.0773	0.0320	-0.0674	0.0298	-0.0606	0.0296
	edu	0.0163	0.0094	0.0113	0.0071	0.0071	0.0050	0.0038	0.0029
	child	-0.0195	0.0085	-0.0136	0.0063	-0.0084	0.0042	-0.0041	0.0022
	hwh	0.0697	0.0509	0.0612	0.0470	0.0482	0.0392	0.0259	0.0208
	_cons	-0.1313	0.1226	-0.0714	0.0734	-0.0262	0.0365	-0.0027	0.0126
MLE	age	0.6611	0.2041	0.6353	0.1706	0.5957	0.1279	0.5440	0.0882
	age2	-0.1272	0.0296	-0.1235	0.0255	-0.1176	0.0205	-0.1096	0.0161
	edu	0.0366	0.0029	0.0279	0.0021	0.0193	0.0014	0.0103	0.0007
	child	-0.0308	0.0070	-0.0229	0.0052	-0.0151	0.0034	-0.0074	0.0017
	hwh	0.1716	0.0274	0.1567	0.0259	0.1246	0.0226	0.0605	0.0149
	cons	-0.2392	0.1368	-0.1622	0.0847	-0.0888	0.0408	-0.0291	0.0125

Table 2. Fuzzy Wage Equation

Model	Variables	Fuzzy Wage Equation							
		α = 0.2		α = 0.4		α = 0.6		α = 0.8	
		Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.
Two-Step	edu	-0.0097	0.0086	-0.0060	0.0057	-0.0038	0.0033	-0.0021	0.0015
	pexp	0.0459	0.1182	0.0402	0.1164	0.0349	0.1157	0.0315	0.1165
	pexp2	-1.1815	3.1363	-1.4309	2.9672	-1.4383	2.8643	-1.1739	2.8415
	pexpchd	0.0187	0.0340	0.0137	0.0314	0.0118	0.0293	0.0140	0.0283
	pexpchd2	-0.0063	0.0111	-0.0046	0.0103	-0.0040	0.0096	-0.0046	0.0093
	_cons	0.0731	0.0684	0.0677	0.0448	0.0487	0.0253	0.0228	0.0109
MLE	edu	0.0001	0.0040	0.6353	0.1706	0.0000	0.0019	0.0000	0.0009
	pexp	0.0586	0.1119	-0.1235	0.0255	0.0590	0.1110	0.0747	0.1102
	pexp2	-2.8597	2.6546	0.0279	0.0021	-2.5953	2.6501	-2.4354	2.6599
	pexpchd	-0.0038	0.0268	-0.0229	0.0052	-0.0016	0.0265	0.0028	0.0263
	pexpchd2	0.0009	0.0088	0.1567	0.0259	0.0004	0.0087	-0.0010	0.0086
	_cons	0.1226	0.0547	-0.1622	0.0847	0.0546	0.0239	0.0199	0.0103

8. Conclussion and Discussion

Heckman's two step estimator is often used in various fields of economics such as labour economics. However, Heckman's two-step estimator often performs poorly. It is necessary to use the maximum likelihood method to estimate the models in such cases. Fuzzy Parametric Sample Selection Model (FPSSM) used to solve the problem of uncertainty of the data obtained using the Parametric Sample Selection Model (PSSM).

REFERENCES

1. Amemiya, T., Tobit Models: A Survey, *Journal of Econometrics*, 1984, 24: 3-61.
2. Amemiya, T., *Advanced Econometrics*, Harvard University Press, 1985.
3. Heckman, J.J., Shadow Prices, Market Wages and Labor Supply, *Econometrica*, 1974, 42: 679-694.
4. Heckman, J.J., The Common Structure of Statistical Models of Truncation, Sample Selection Bias and Limited Dependent Variables and A Simple Estimator for Such Models, *Ann. Econom. Social Measurement*, 1976, 5: 475-492.
5. Heckman, J.J., Sample Selection Bias as A Specification Error, *Econometrica*, 1979, 47: 153-161.
6. Lola, M.S., Kamil, A.A., and Abu Osman, M.T., Fuzzy Parametric of Sample Selection Model Using Heckman Two-Step Estimation Models, *American Journal of Applied Sciences*, 2009, 6(10): 1845-1853.
7. Maddala, G.S., *Limited-Dependent and Qualitative in Econometrics*. Cambridge: Cambridge University Press, 1983: 257-289.
8. Mroz, T.A., The Sensitivity of an Empirical Model of Married Women's Hours of Work to Econometric and Statistical Assumptions, *Econometrica*, 1987, 55: 765-799.
9. Nawata, K., *Estimation of Sample Selection Bias Models*, 1996: 387-400.
10. Newey, W., *Two Step Series Estimation of Sample Selection Models*, Department of Economic, MIT working paper, 1988, No. E52 - 262D: 1-17.
11. Zadeh, L.A., *On The Analysis of Large-Scale System.* In: Gottinger, H., ed., Systems approaches and Environment Problems. Vandenhoeck and Ruprecht, 1974.