Fuzzy Semi-parametric Sample Selection Model Case Study for Participation of Married Women

L. MUHAMAD SAFIIH [1]*, A.A.BASAH KAMIL[2], M. T. ABU OSMAN [3]
[1, 3]Mathematics Department
Faculty of Science and Technology, University Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, MALAYSIA.
safiihmd@umt.edu.my, abuosman@umt.edu.my, http://www.umt.edu.my

[2]School of Distance Learning
Universiti Sains Malaysia
11800 USM Penang, MALAYSIA.
anton@usm.my; http://www.usm.my

Abstract: The sample selection model is studied in the context of semi-parametric methods. The issue of uncertainty and ambiguity are still major problems and the modelling of a semi-parametric sample selection model as well as its parametric. The best approach of accounting for uncertainty and ambiguity is to take advantage of the tools provided by the theory of fuzzy sets. The semi-parametric of a sample selection model is an econometric model that has found an interesting application in empirical studies. In this paper, the married women participants in the Malaysia labour force are studied. It comprises the analysis of a) participation equation in the wage sector and b) the wage equation in the wage sector. The data set used for this study is from the Malaysian population and family survey 1994 (MPFS-1994).

Key-words:- uncertainty, semi-parametric sample selection model, participant equation, wage equation, fuzzy number.

1 Introduction

The study of semi-parametric sample selection models has received considerable attention from statisticians as well as econometricians in the late 20th century (see Schafgans, 1996). The termed "semi-parametric," used as a hybrid model for the selection models, which do not involve parametric forms on error distributions; hence, only the regression function part of the model of interest is used. Consideration is based on two perspectives, firstly; no restriction of estimation of the parameters of interest for the distribution function of the error terms, secondly; restricting the functional form of heteroskedasticity to lie in a finite-dimensional parametric family (Schafgans,1996). Gallant and Nychka (1987) studied these methods in the context

of semi-nonparametric maximum likelihood estimation and applied the method to nonlinear regression with sample selection model. Newey (1988) used a series approximation to the selection correction term that considered regression s-pline and power series approximations. Robinson (1988) focused on the simplest interesting setting of multiple regressions with independent observations, and described extensions to other econometric models, in particular seemingly unrelated and nonlinear regressions, simultaneous equations, distribution lags and sample selectivity models. More specifically, none of these researchers put any efforts into studies that analysed semi-parametric sample selection models in the context of fuzzy environment like fuzzy sets, fuzzy logic or fuzzy sets and systems (Muhamad Safiih, 2007). Using *S*-curve membership function, Muhamad Safiih *et al.*(2006), introduce a concept of membership function to the sample selection model. A real data set especially income from Malaysian Family and Population Survey 1994 used to get the upper and lower membership function and compared to the crisp data.

However, the concept of fuzzy membership function have been widely use in other area. The first article based on fuzzy linear regression model was by Tanaka, Hayashi and Asai (1982). Through a regression problem, fuzzy dependent variable and crisp independent variable formulated as a mathematical programming problem. Kao and Chyu (2002) used this concept on regression coefficients that provided the best explanation for the relationship between the independent and dependent variables.

This paper introduces a membership function of a semi-parametric sample selection model in which historical data contains some uncertainty. The concept of fuzzy sets by Zadeh is extended from the crisp sets, that is the two-valued evaluation of 0 or 1, {0, 1}, to the infinite number of values from 0 to 1, [0, 1]. (see Terano *et al*. 1994). This provides an ideal framework to deal with problems in which there does not exist a definite criterion for discovering what elements belongs or does not belong to a given set (Miceli, 1998). Fuzzy set are defined by a fuzzy sets in a universe of discourse *U* is characterised by a membership function denoted by the function μ_A maps all elements of *U* that take the values in the interval [0,1] that is $A: X \rightarrow [0,1]$ (Zadeh, 1965).

2 Data description

2.1 Data description and Variables used

2.1.1 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 the National Population and Family Development Board of Malaysia under the Ministry of Women, Family and Community Development of Malaysia. This survey was specifically for married women and it provides relevant and significant information such as wages, educational attainment, household composition and other socioeconomic characteristics.

The original data set consisted of 4444 married women. Following the sequential criteria by Mroz (1984), only married women with completed information are considered. For those with incomplete the information removed from the sample. The resulting sample data set contains 2792 women. To create the sample of women on the basis of the MPFL-94 data set used the selection rules of Martins (2001). As a result left only 1100 (39.4 %) of sample data set whom were employed and considered as participants. The rest were considered as non-participants i.e. 1692 married women (60.6% unemployed). This study consisted of 2792 married women (both participant and non-participant) data sets. The descriptive statistics of the data set is representation in Table 1. To create the sample criteria choosing for participant and non-participant married women based on the MPFS-94 data set, are as follows:

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

2.1.2 Variables Used in the Study

In this study following the literature (see Gerfin, 1996; Martins, 2001; Christofides, Li, Liu, and Min, 2003), fuzzy semi-parametric of a sample selection model consists of two equations or parts.

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

The first equation is a participation equation i.e. the probability of married women participating in the labour market. The 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 18 living in the family), and *HW* (log of monthly husband's wage). The standard human capital approach is followed for the determination of wages, except for 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 is adopted using a method advanced by Buchinsky (1998). The participation equation is given by

$$d_i^* = \beta_0 + \beta_1 AGE + \beta_2 AGE2 + \beta_3 EDU + \beta_4 CHILD + \beta_5 HW + u_i \quad (2)$$

where;

$$Participate = 1(d_i^* > 0)$$

Consider 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 that was included with Q_z is given by:

$$Q_z = \gamma_1 PEXP + \gamma_2 PEXP2 + \gamma_3 PEXPCHD + \gamma_4 PEXPCHD2 \quad (3)$$

The second equation is called outcome equation. The dependent variable used for the analysis was the log hourly wages (z). The independent variables were *EDU* (education), *PEXP* (potential work experience divided by 10), *PEXP2* (potential experience squared divided by 100), *PEXPCHD* (*PEXP* interacted with the total number of children) and *PEXPCHD2* (*PEXP2* interacted with the total number of children). Both participation and wage equation were considered as specification I and II respectively i.e. the most basic one of SSM.
When *Participate* = 1, the outcome equation is observed and given by

$$In(Y) = \gamma_0 + \gamma_1 EDU + \gamma_2 PEXP + \gamma_3 PEXP2 + \gamma_4 PEXPCHD + \gamma_5 PEXPCHD2 + \varepsilon_i \quad (4)$$

According to Kao and Chyu (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 the observations fall into the category of fuzzy regression analysis. Kao and Chyu (2002) also highlighted that if the data are fixed it is called non-fuzzy data, i.e., in terms of integer value, the data cannot be fuzzified, and then it is considered fuzzy data. For instance, in this study *are EDU* and *CHILD*. Assumed data used in this study contained uncertainty, instead of crisp data, therefore fuzzy data are more appropriate. In the participation equation, fuzzy data used for the independent variables (x) involves *AGE* (age in year divided by 10), *AGE2* (age square divided by 100) and *HW* (log of monthly husband's wage). For the outcome equation, the fuzzy data that was used for dependent variable was the log hourly wages (z) while the independent variables (x), fuzzy data involved the variables *PEXP* (potential work experience divided by 10), *PEXP2* (potential experience squared divided by 100), *PEXPCHD* (*PEXP* interacted with the total number of children) and *PEXPCHD2* (*PEXP2* interacted with the total number of children).

2.1.2.1 Endogenous Variables

In this study "participation equation" was the first dependent variable. This variable is dichotomous indicator with the value 1 if the women participate and 0 otherwise. 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 other races were 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 than basic wages, for instance allowance, bonus, etc, are an important part of total earning (Mazumdar, 1991). Schafgans (1996). Therefore, bonuses and payments in-kind (for instance food, housing, etc) are included in the computational of hourly wages. From the 1994 survey, comparison of the wages sector especially labour market, the Chinese women has significantly higher income wages (≥ RM3, 000.00 or 1.1%) than for Malay and Indian income wages (0.9%). For income less then RM999.00, the lower hourly wages for married women are close for Malay (96.1%), Chinese (94.1%) and Indian (96.3%). Schafgans (1996) divided the exogenous variables into two groups. The first group consists of variables which influence the reservation wage and do not influence the offered wage. The second group consists of variables which potentially influence both the offered

and reservation wage. The variables listed in the first group will be used to identify the decision to participate in the wage sector from the wage determination.

2.1.2.2 Exogenous Variables

In this second group, the variables listed are exogenous variables which enter both in the participation equation and in the wage equation. AGE and EDU (education) are involved in the participation equation, whereas, EDU and potential experiences are involved in the outcome equation. The purpose of using EDU is to measure general human capital and they are expected to have negative effect on the probability of being employed.

AGE: The 1994 survey shows that the age of women wage workers (on average) is 34.45 years old while the age for women non-wage workers is 35.7 years old. These indicate that the women participating in the labour market are younger than for non-participating women. This result is contradicts with Schafgans (1996) i.e. it is not consistent with the increased importance of the wage sector in Malaysia. As result, the younger women participant (on average) is 1.61 against 2.07 in labour market, which means the women participating in the labour market are less educated. The age variable is used to measure general human capital and is expected to have negative effect on the probability of being employed.

Education: The education attainment is measured by the continuous variable i.e. "*years of schooling*". This variable measures the years of schooling required to obtain the highest grade completed. No measure was available regarding the actual years it took for each individual to reach the level completed (Schafgans, 1996). 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 (see Appendix B). The details of schooling for the Malaysian education system can be found in Schafgans (1996).

The potential experience: This was calculated by $age_i - schooling_i - 6$ with women participants (15.4 years) less compared to women non-participants (20.8 years). This implies that the women participants in the labour market are influenced or is likely to be interrupted by her childbearing and child-raising activities. Data from MPFS-94 shows that the total numbers of children (under 7 years old) for women non-participant are higher than for women participant i.e. 965 children versus 441 children. Although the total fertility rate in Malaysia decreased to 4.0% (1985-1990), 3.6% (1990-1995) and 3.2% (1995-2000) compared to 6.3% in 1965 (United Nations, 2001).

The data set used for this study is from the Malaysian population and family survey 1994 (MPFS-1994). The original MPFS-94 sample data comprises 4444 married women. Based on the sequential criteria (Mroz,1984), this analysis was limited to the completed information provided by the married women. The resulting sample data set consisted of 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 total 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 the participant and non-participant married women based on the MPFS-94 data set.

3 Empirical results: semi-parametric and fuzzy semi-parametric SSM

This section presents the results that applies to the most basic one i.e. the participant and wage equation of DWADE estimator (Hardle *et al*. 1995) and Powell estimator (Powell,1987) respectively. Both estimators are consistent with $\sqrt{n}-$consistency and asymptotic normality. The discussion focuses on the participation and wage equation in terms of the estimated coefficient, the significant effect and consistency of the estimate for SPSSM, as well as FSPSSM for comparison purposes.

3.1 Participation equation in the wage sector

The participation equation using DWADE estimator and FSPSSM estimator results are presented in Table 2. The first column by DWADE estimator with value of bandwidth h = 0.2. 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. The values in the parenthesis are the standard error of mean. From Table 1, the coefficient

estimate for participation equation of SPSSM shows a positive coefficient for *EDU* and *HW*, negative coefficient for *AGE*, *AGE2* and *CHILD*. However, only variables *EDU* and *CHILD* show significance at 5% level. The results suggest that *EDU* and *CHILD* (the number of children in the family) are important factors in a women's participation decision.

Applying the FSPSSM, the coefficient estimate shows a similar trend with SPSSM i.e. a positive coefficient for variables *EDU* and *HW*, negative coefficient for variables *AGE*, *AGE2* and *CHILD*. A significant at 5% level for variables *EDU* and *CHILD*. The coefficient estimated obtained from FSPSSM was considerably quite close to the coefficient estimated by conventional SPSSM. In term of efficiency, almost all the standard errors for the variables of FSPSSM (except for variable *AGE2* with α - cuts of 0.0 0.2 and 0.04) were much smaller (0.09 % - 8.3%) compared to those in the conventional SPSSM. In other words, for participation equation of FPSSM, the coefficient estimates are relatively close to SPSSM as the values of α-cuts increased (from 0.0 to 0.8) and almost smaller of standard error than PSSM.

3.2 The wage equation in the wage sector

The wage equation of the Powell estimator and FSPSSM results are presented in Table 3. The first column used Powell estimator with values of bandwidth h = 0.2 without the constant terms. The following columns are results given by fuzzy semi-parametric sample selection model (FPSSM) with $\alpha-$cuts 0.0, 0.2, 0.4, 0.6 and 0.8 respectively. The values in the parenthesis are the standard error of mean.

From Table 3, estimating the coefficient for wage equation of SPSSM shows a positive coefficient for *PEXP* and *PEXPCHD2*, negative coefficient *EDU*, *PEXP2* and *PEXPCHD*. All variables show significance at 5% level. The results suggest that all variables are influence factors in a women's wage. Applying the FSPSSM, estimating the coefficient shows a similar trend with SPSSM i.e. positive coefficient for *PEXP* and *PEXPCHD2*, negative coefficient for *EDU*, *PEXP2* and *PEXPCHD*. All variables are significant at 5% level. The coefficient estimate obtained from FSPSSM also was considerably close to the coefficient estimated by conventional SPSSM. In other words, applying FSPSSM, estimating the coefficient is consistent even though it involves uncertainties in the data. In term of efficiency, almost all the variables (except for variables *PEXP2* and *PEXPCHD* with α - cuts of 0.0, 0.2, 0.04 and 0.06) show equality or were quite close to standard error (0.003 % - 0.09%) as compared to those in the conventional SPSSM. That means, both methods are considerably equal in term of efficiency. This evidence shows that FSPSSM is considerably efficient as in the conventional semi-parametric model.

4 Conclusion

For comparison of the participant equation, the estimated coefficient and the significant factor gives a similar trend with the SPSSM. However, one of the interesting findings and the most significant result appears by applying the FPSSM i.e. the FSPSSM is a better estimate when compared to the SPSSM in terms of the standard error of the coefficient estimate. The standard errors of coefficient estimate for the FSPSSM almost give a smaller when compared to the conventional SPSSM. This is evidence that this approach is much better in estimating coefficient and has a considerable efficiency gain then those in the conventional semi-parametric model. The coefficient estimated obtained was also considerably closer with the coefficient estimated by conventional SPSSM. Hence, this gives evidence that the coefficient estimated is consistent although data used involved uncertainties. Secondly, wages equation, in terms of the coefficient estimation and significant factor, the FSPSSM is considerably closer to the standard error of SPSSM. As a whole, the FSPSSM gave a better estimate when compared to the SPSSM. In terms of consistency, it was found that the coefficient estimate for all variables of FSPSSM were not much different to the coefficient estimate of SPSSM, even though the values of the $\alpha-cuts$ increased (from 0.0 to 0.8). In the other words, by looking at the coefficient estimate and consistency, the fuzzy model (FSPSSM) is much better than the model without fuzzy (SPSSM) for the participation and wage equation.

References:

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

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

[3] Gallant, R. and D. Nychka (1987). Semi-Nonparametric Maximum Likelihood estimation, Econometrica, 55, p. 363-390.

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

[5] Härdle, W. Klinke, S. and Müller, M. (1999). Xplore learning guide. Berlin. Springer-Verlag.

[6] Kao, C and Chyu, C-L. (2002). A Fuzzy regression model with better explanatory power. Fuzzy Sets and Systems 126. p. 401-409.

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

[8] L. Muhamad Safiih, A.A. Basah Kamil and M.T. Abu Osman (2006). Fuzzy Approach to Sample Selection Model. WSEAS Transactions on Mathematics. Issue 6. Volume 5. p. 706-712.

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

[10] Mazumdar, D. (1991). Malaysian labor markets under structural adjustment. Pre Working paper No.573. Washington D.C.: World Bank.

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

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

[13] Newey, W. (1988). Two step series estimation of sample selection models, Department of Economic, MIT working paper no. E52 - 262D. p. 1-17.

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

[15] Robinson, P.M. (1988). Root-N consistent semiparametric regression, Econometrica, 56, p.931-954.

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

[17] Tanaka, H., Hayashi, S. and Asai, K. (1982). Linear regression analysis with fuzzy model. IEEE Transactions on Systems, Man and Cybernetics, 12. p. 903-907.

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

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

Table 1: *Descriptive statistics*

Variables	Participants (n=1100)		Non-participants (n=1692)		Total (n=2792)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
AGE	34.45	7.482	35.7	7.722	35.21	7.651
AGE2	12.43	5.21	13.34	5.43	12.98	5.364
EDU	1.61	1.289	2.07	1.222	2.12	1.230
CHILD	8.65	4.390	7.00	3.930	7.65	4.194
HW	2.062	1.05	2.46	0.87	2.65	0.584
Y	1.58	0.325	-	-	-	-

Definition of the variables

Z is the indicator of labor market participation

AGE (in years) is married women's age divided by 10

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

HW-S is the log of the husband's monthly wage (measure in Ringgit Malaysia)

Y is the log women's hourly wage rate (measured in Ringgit Malysia)

PEXP is potential work experience, defined as age minus years of schooling

PEXP2 is potential work experience squared

PEXPCHD is potential work experience times the number of children

PEXPCHD2 is potential work experience squared, times the number of children

Table 2: Semi-parametric and fuzzy semi-parametric estimates for the participation equation

Participation equation	***Coefficients***					
	DWADE	***fuzzy selection model***				
		α ***= 0.0***	α ***= 0.2***	α ***= 0.4***	α ***= 0.6***	α ***= 0.8***
AGE	-0.002048	-0.001642	-0.0016184	-0.0015934	-0.0043978	-0.0015393
	(1.233)	(**1.232**)	(**1.232**)	(**1.230**)	(**1.151**)	(**1.150**)
AGE2	-0.00016099	-0.00016673	-0.00016651	-0.00016629	-0.00020722	-0.00016584
	((0.1754)	(0.1767)	(0.1765)	(0.1763)	(**0.1627**)	(**0.1624**)
EDU	0.00034766*	0.00023044*	0.00023044*	0.00023044*	0.00011323*	0.00023044*
	(0.02116)	(**0.02062**)	(**0.02062**)	(**0.02115**)	(**0.02015**)	(**0.02015**)
CHILD	-0.0039216*	-0.0044301*	-0.0044301*	-0.0044301*	-0.0048986*	-0.0044301*
	(0.06573)	(**0.06484**)	(**0.06485**)	(**0.06571**)	(**0.0634**)	(**0.06341**)
HW	0.044008	0.048832	0.049189	0.049549	0.05597	0.050262
	(0.1632)	(**0.1432**)	(**0.1437**)	(**0.1485**)	(**0.1396**)	(**0.1402**)

** 5% level of significant*

Table 3: Semi-parametric and fuzzy semi-parametric estimates for the wage equation

Wage equation	**Coefficients**					
	Powell	**fuzzy selection model**				
		α ***= 0.0***	α ***= 0.2***	α ***= 0.4***	α ***= 0.6***	α ***= 0.8***
EDU	-0.0112792*	-0.0114256*	-0.011385*	-0.011346*	-0.010939*	-0.0109003*
	(-0.005262)	**(-0.005258)**	**(-0.005259)**	**(-0.005259)**	**(-0.005258)**	**(-0.005258)**
PEXP	0.544083*	0.530069*	0.532247*	0.534385*	0.538776*	0.540864*
	(-0.1099)	**(-0.109)**	**(-0.1092)**	**(-0.1093)**	**(-0.1094)**	**(-0.1096)**
PEXP2	-0.160272*	-0.158259*	-0.158525*	-0.158781*	-0.159524*	-0.159762*
	(-0.02633)	**(-0.02632)**	**(-0.02632)**	**(-0.02632)**	**(-0.0263)**	**(-0.0263)**
PEXPCHD	-0.161205*	-0.158262*	-0.158584*	-0.15889*	-0.159583*	-0.159863*
	(-0.02453)	(-0.02463)	(-0.02461)	(-0.02459)	(-0.02455)	**(-0.02453)**
PEXPCHD2	0.046591*	0.0455835*	0.0457004*	0.0458118*	0.0462221*	0.0463242*
	(-0.008485)	(-0.008517)	(-0.008511)	(-0.008508)	(-0.008493)	**(-0.008485)**

**5% level of significant*