Journal of Rural Development 18(Summer 1995) : 17~27

A SEMIPARAMETRIC ANALYSIS OF FARMERS' CHOICE ON THE OFF-FARM TRAINING

KYEONG-DUK KIM*
JI-HYEON CHOI**

I. Introduction

In the traditional Parametric approach of estimation, OLS, for instance, it has been recognized that the distribution of error terms only affects the estimators' efficiency by the facts that the distribution form of error terms could not affect estimators' efficiency and that as sample size increases, estimators have a normal distribution form (Central Limit Theorem). Thus, studies on non-stochastic factors, such as functional forms of estimation equations or exogenous test methods, have been actively performed while those on the distributions of error terms have been rarely done.

However, in discrete choice models, the estimators are not consistent if the assumed distribution of the error terms are different from their real distributions (Manski 1975; Cosslett 1983). Whether or not the estimators are consistent depends on the validity of assumptions on the distributions of error terms. Therefore, profit or logit models can generate the most consistent and efficient estimators if the assumptions on the specific distribution forms of error terms, i.e., normal or logistic distribution forms, are correct. But the estimators by probit or logit method cannot meet the consistency criteria (even if they can meet the efficiency criteria) if the assumptions on the distribution forms of the error terms are not correct. That is, the assumption of the distribution on error terms is

* Research Associate, Korea Rural Economic Institute.
** Fellow, Korea Rural Economic Institute, Authors would like to thank two reviewers for valuable comments and suggestions.

essential for consistent estimates.

To avoid these problems, methods that do not require specifying the distribution of error terms have been introduced. Such methods are called semiparametric because they involve an unknown distribution of error terms as well as the unknown finite-dimensional parametric vector of regression equations.

Since late 1970's, several semiparametric econometric methods have been suggested. Despite this development, application of the methods does not prevail.[1] In this paper, the Maximum Score method proposed by Manski(1975 and 1985) is used to analyze farmer's choice for the off-farm training in Korea.

In Korea, agriculture reforms such as large-scale farming and off-farm income improvement are important for agricultural market opening. Land mobilization is required for successful large-scale farming, it requires small-scaled farmers to change or quit their current farm jobs. Therefore, off-farm training programs are needed to encourage small-scaled farmers to change their current farm activities into off-farm activities.

There are many studies on the farmers' off-farm job choices (Tolbert 1974; Norris et al. 1979). However, only a few focus on the farmers' choice analysis for the off-farm training programs except Lee (1992). Lee analyzed the occupational choice pattern of the farm household members. He found many important variables to affect farm household members' attitude toward the off-farm training program such as the age of the farm head, education level of the head, and farm size. But he used the *OLS* estimation technique although the dependent variable is discrete. The estimation technique consequently yields inconsistent estimates of the parameters (Wales and Woodland 1983).

In relation to the issues raised above, the economic and socio-demographic aspects affecting farmers' choice on the off-farm training programs are of interest in this research. The objective of this study is to assess the efficiency and consistency criteria when a semiparametric method and the usual parametric method, such as probit and logit estimation are applied.

---

[1] Exceptions are Horowitz and Neumann(1987), Newey, Powell, and Walker(1990), and Nahm and Lees(1993).

This paper consists of five sections. The section 2 describes the semiparametric maximum score estimation method. The section 3 explains the data set and reports the estimated results. The section 4 derives policy implications from estimated results. Finally, concluding remarks are given in the section 5.

II. Estimation Methods

Consider a binary choice model in the following equation:

$$y = l(\beta' x + \varepsilon > 0) \tag{1}$$

Where $l(A)$ is an indicator function of the event A. It is one if the dependent variable $y^*(= \beta' x + \varepsilon)$, generated by the regressor x and unobserved error term ε , is greater than 0 and zero otherwise. Then the above equation can be rewritten as follows:

$$Pr(y = 1 \mid x) = 1 - F(-\beta' x) = F(\beta' x) \tag{2}$$

$$Pr(y = 0 \mid x) = 1 - F(\beta' x) \tag{3}$$

Where $F(\cdot)$ is the cumulative probability function of error terms.

The parametric approach typically uses maximum likelihood methods to estimate β assuming a specific distribution for error terms. We have a probit or logit model, according to the assumption on the distribution of error terms. if the distribution follows a normal (logistic) distribution, the probit (the logit) model provides the best unbiased and efficient estimates under the standard assumption on the errors. However, if the assumption about the error distribution is not consistent with true distribution, maximum likelihood estimation methods cannot assure even consistency.

In this paper, the maximum score estimation method proposed by Manski(1975 and 1985) is applied. The maximum score estimation (MSE) method only assumes the (unobserved) distribution of errors conditioned by explanatory variables, $F(\cdot)_{u|x}$, which is assumed to have a zero median for consistent estimators. That is, when the

median of the error term is zero, MSE method has the consistent estimator, $\beta_* = \beta / \| \beta \|$ (where the notation $\| \|$ expresses the norm), which is not the case for the logit or probit estimates (Manski 1985). Therefore, we can use the MSE method for consistent estimates even if the specific functional form is unknown and the error distribution is heteroskedastistic.

We can get the consistent MSE estimator by maximizing the following sample score functions.

$$arg \max S_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} [z_i \, \mathrm{sgn}(x_i \beta)], \quad i = 1, 2, \ldots, n \tag{4}$$

Where n is the sample size and $S_n(\beta)$ is a score function. $z = 2y - 1$ and $\mathrm{sgn}(\cdot)$ are defined as follows:

$$\begin{aligned} &z = 1, \quad if \quad y^* = \beta' x + \varepsilon > 0 \\ &z = -1, \quad if \quad y^* = \beta' x + \varepsilon \leq 0 \end{aligned} \tag{5}$$

$$\begin{aligned} &\mathrm{sgn}(d) = 1, \quad if \quad d > 0 \\ &\mathrm{sgn}(d) = -1, \quad if \quad d \leq 0 \end{aligned} \tag{6}$$

It turns out that maximization of $S_n(\beta)$ yields a consistent estimate of β. Note that the MSE estimator only identifies β up to scale.

Intuitively, the maximization of $S_n(\beta)$ is to maximize the number of correct predictions. For given β, if the $\mathrm{sgn}(x_i \beta)$ is equal to the observed responses z_i, 1/n is added to the sample score function as $x_i \beta$ predicts z_i correctly. If they are not the same, then 1/n is subtracted from the sample score function as $x_i \beta$ predicts incorrectly. Through this process, the value of β maximizing the number of correct prdictions is the maximum score estimates.

From the above, we know that the MSE method is different form the conventional maximum likelihood estimation methods that maximize certain functions. The sample score function is not a continuous function but a step function, and the derivatives of the step function do not provide information to choose adequate search directions from the initial estimation. This makes the MSE method

different from the conventional methods.

The MSE method does not choose the most adequate search direction. It chooses k-1 orthogonal direction sets and maximizes the sample score by iteration according to the selected directions (Manski and Thompson 1986). If the score after one more iteration is the same as the previous iteration score, we consider the sample score function converged to a maximum value. But we cannot assure this value is the global maximum. That is, there is some probability for this value to be locally maximized. So we must take an end-game searching. Each end-game searching does iteration with a new orthogonal search-direction sets selected arbitrarily.

However, there is no sampling or asymtotic distribution theory for the MSE method yet. Thus, the standard errors of the MSE have been obtained by bootstrapping. The accuracy of bootstrap estimates has been evaluated by Manski and Thompson(1986) by means of Monte Carlo experiments. Their results suggest that the bootstrap provides enough information about the precision of the MSE to make the bootstrap useful in empirical research. They further suggest a more conservative approach that doubles the bootstrap estimate of the maximum score standard error and takes this as an upper bound on its true value.

III. Data and Estimated Results

The farmers' choice on the off-farm training is of interest in this study. The training period is longer than or equal to 6 months. The data are from the National Survey on farmers' management of the Office of Rural Development and field survey on farmers' attitude on the domestic agricultural market opening by the Korea Rural Economic Institute in 1994, respectively. The sample data consists of 124 farmers living in Jungsun-Gun, Namjejoo-Gun, and Kimje-Gun.

The theoretical background of the training decision is provided by Ehrenberg and Smith(1988). They refer to education, training, migration, and search for new jobs as investment in human capital. Ehrenberg and Smith suggested age, opportunity cost, and earning differentials could be incorporated with the demand for education or training. The human capital model of mobility also suggested that the

level of wages and region as well as age and education, which are the personal characteristics of movers, are important factors influencing employer-initiated mobility.[2]

Based on the theoretical framework of Ehrenberg and Smith, the independent variables for the model include the age of the farm head, an education level of the head, agricultural income, non-agricultural income, the size of cultivated land, the number of family members, farm household debt, and a regional dummy variable.

We include the regional dummy variable because there is a large opportunity for off-farm employment in Kimje than other regions. Thus, farmers in Kimje are more likely to take off-farm training. It is hypothesized that the farmers' behavior living in Kimje province is different from that of the other's. The estimated model is specified as follows:

$$\begin{aligned} y = 1 \quad & if \ \beta_0 + \beta_1 AGE + \beta_2 EDU + \beta_3 FINCM + \beta_4 OFFINCM \\ & + \beta_5 LAND + \beta_6 FAM + \beta_7 DEBT + \beta_8 REG + \varepsilon \qquad (7) \\ = 0 \quad & \text{otherwise,} \end{aligned}$$

where AGE = age of head
EDU = education level of head
$FINCM$ = agricultural income
$OFFINCM$ = non-agricultural income(exempted transferred income)
$LAND$ = size of cultivated land
FAM = number of family members
$DEBT$ = farm household debt for agricultural production
REG = 1 if residence is in Kimje-Gun, and 0 otherwise
y = 1 if accepting off-farm training, and 0 otherwise

The estimated results and prediction results are given in Table 1 and Table 2, respectively. The probit and logit model predict 101 of 124 or 81.5%, while the MSE model predicts 103 of 124, or 83.1% of the observations correctly. Therefore, we can conclude that for the case where the distribution of error terms is not known to us correctly, the MSE method gives us better prediction than the probit or logit

[2] Refer to Ehrenberg and Smith(1988) for the detailed discussion of the migration model.

TABLE 1 Estimated Coefficients and T-Values

Variables	PROBIT	LOGIT	MSCORE
constant	0.84871	1.2646	-0.14344
	(0.744)	(0.650)	(-0.337)
AGE	-0.04044	-0.065401	-0.14864
	(-2.298)	(-2.184)	(-1.952)
EDU	-0.07785	-0.13494	0.078041
	(-0.509)	(-0.513)	(0.234)
FINCM	0.26694E-04	0.52486E-04	0.21823
	(1.471)	(1.600)	(0.815)
OFFINCM	0.22319E-05	0.38348E-05	-0.40131
	(0.124)	(0.119)	(-1.261)
LAND	-0.50551E-04	-0.96899E-04	-0.50287
	(-1.916)	(-1.916)	(-1.873)
FAM	0.012213	0.22015E-04	-0.47937
	(0.106)	(0.110)	(-0.874)
DEBT	-0.25995E-05	-0.51995E-05	-0.47742
	(-2.223)	(-2.253)	(-1.534)
REG	1.1075	1.8814	1.17875
	(3.3296)	(3.247)	(2.397)

TABLE 2 Model Performance on Prediction

		PROBIT		LOGIT		MSCORE		
		Predicted		Predicted		Predicted		
		0	1	0	1	0	1	Total
Actual	0	96	3	96	3	96	3	99
	1	20	5	20	5	18	7	25
Total		116	8	116	8	114	10	124

model for the binary choice model. That is, when the assumption about the distributions of the error terms is not correct, it is more conservative to use the MSE model.

IV. Some Implications for the Estimated Results

First, in binary choice models, estimated results are different according to the methods used to estimate the coefficients. Results from the MSE method are less efficient than those from the probit or logit. That is, for the MSE method, t-values are lower. This is because the MSE method does not assume a specific distribution of the error terms.[3] And coefficients for the education level of the head, non-agricultural income and number of farm family members have different signs across the estimation methods.

However, the coefficients for *AGE*, *LAND* and *REG* that are estimated by MSE method are significant as in the probit or logit results. So we conclude that the key variables affecting the decisions on the off-farm training, which is continued for 6 months or more, the age of head, the size of cultivated land and living location.

Second, we infer that the higher the age of the head and the larger the land size, the lesser the chance of farm heads ' participation in the off-farm training. These can be explained by the fact that as farmers get older, it is more difficult for the old to take off-farm jobs, and with larger cultivated lands, there is a higher opportunity cost to change their farming job into off-farming jobs.

As an interesting point, the coefficient of the region dummy variable is very significant. This result implies that in the plain regions like Kimje-Gun, there is a larger opportunity for farmers to get off-farm jobs without moving to other locations that provide off-farm jobs. This is mainly because of transportation convenience with the neighboring regions.[4] And even though the competitive advantage

---

[3] It is well known that if the distribution of the error terms are known correctly, using maximum likelihood estimation methods generates best unbiased estimators. Furthermore, t-values of the MSE method is not accurately calculated but approximated with the bootstrap mean-squared deviations.

[4] There are some cities, for example, Jonju, Iree, and Kunsan, which contains more non-agricultural activities. This cities are near to Kimje-Gun, it takes only one hour by bus.

agricultural activities are cropping fields with rice in the plain regions for the given conditions, income from cultivating rice are less than that from other grains. Therefore, it is natural for farmers living in the plain regions to more likely participate in off-farm training programs.

The coefficient of farm household debts from the probit and logit model has a minus sign and is significant while it is insignificant in the MSE model. This implies that the more farm household debt, the less inclination to take the off-farm training program. It can be interpreted such that farm heads who have more debt for agricultural production are afraid of repaying their debts or being treated with unfavorable conditions when they have full-time off-farm jobs instead of farming. So we can conclude that the success of the off-farm training program highly depends on how to handle the existing farm household debts.

Third, coefficients of education level of the head, non-agricultural income and the number of farm family members are insignificant and those coefficients have different signs across the estimation methods. It was expected that as education level of farm head is higher, the heads have more intention to participate in the off-farm training program because fewer efforts are demanded to get new technologies. Furthermore, higher non-farm income from the existing off-farm jobs and advanced off-farm income from the existing off-farm jobs and advanced off-farm technologies make farm heads less likely to participate in new off-farm training programs. Also it is reasonable to think that as the family gets larger, it could take a higher transaction cost in changing family heads' jobs. The MSE method seems to provide more reasonable results in terms of signs of coefficients.

V . Conclusion

We obtained both parametric and semiparametric estimates. The probit and logit estimates differ from the corresponding maximum score estimates. If the assumed distribution of the error terms is different from the actual distribution forms, the estimators could not meet the consistency criteria. In this paper, in order to compare the properties of estimators across the different estimation methods, the

maximum score estimation method, which does not require to assume a specific distribution form of error terms, was applied.

While the MSE method has some merits, the estimator is less efficient than the estimators by probit or logit. In this case, there is no clear cut which estimation method is better. It could be asserted that if consistency is more important, then, the MSE method deserves attention even if the MSE estimator looses some efficiency criteria. In future research, the MSE estimation method might be applied to different areas and the properties of the estimator could be further examined.

REFERENCES

Cosslet, S. 1983. "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model." *Econometrica* : 765-782.

Ehrenberg, R.G. and R. Smith. 1988. *Modern Labor Economics Theory and Public Policy*. Scott, Foresmann and Company.

Green, W.H. 1993. *Econometrics Analysis*. 2nd. ed., Macmillan Publishing Co.

Horowitz, J. L. and G. Neumann. 1987. "Semiparametric Estimation of Employment Duration Models." *Econometric Reviews*: 1-44.

Manski, C. 1975. "Maximum Score Estimation of the stochastic Utility Model of Choice." *Journal of Econometrics* : 206-228.

_____. 1985. "Semiparametric Analysis of Discrete Response: Asymptotic Properties of Maximum Score Estimator." *Journal of Econometrics* 27 : 313-33.

Lee, Y. D.1992. "The Occupational Choice pattern of Farm Household Members and Its Related Variables." Ph.D. Thesis, Seoul National University.

Manski, C. and T. Scott Thompson. 1986. "Operational Characteristics of Maximum Score Estimation." *Journal of Econometrics* 31 : 31-40.

Nahm, J. W. and Y. G. Lee. 1993. "Semiparametric Analysis of Housing Choice Data of Korea." *Sogang Economic Review* 22 : 129-141. Seoul: Sogang University.

Newey, W., J. Powell and J. Walker. 1990. "Semiparametric Estimation of Selection Models: Some Empirical Results."*American Economic Review* 80 : 324-328.

Norris, W. et al. 1979. *The Career Information Service*. Chicago: Rand McNally College Publishing Company.

Tolbert, E. L. 1974. *Counseling for the Career Development*. Houghton : Mifflin

Company.

Wales, T. J. and A. D. Woodland. 1983. "Estimation of Consumer Demand System with Binding Negativity Constraints." *Journal of Econometrics* : 263-285.