Unbiasedness or Statistical Efficiency: Comparison between One-stage Tobit of MLE and Two-step Tobit of OLS

Sun-Young Park

Full-time Lecturer, Department of Venture Technology Management Graduate, School of Venture, Hoseo University

Abstract : This paper tried to construct statistical and econometric models on the basis of economic theory in order to discuss the issue of statistical efficiency and unbiasedness including the sample selection bias correcting problem. Comparative analytical tool were one stage Tobit of Maximum Likelihood estimation and Heckman's two-step Tobit of Ordinary Least Squares. The results showed that the adequacy of model for the analysis on demand and choice, we believe that there is no big difference in explanatory variables between the first selection model and the second linear probability model. Since the Lambda, the self- selectivity correction factor, in the Type II Tobit is not statistically significant, there is no self-selectivity in the Type II Tobit model, indicating that Type I Tobit model would give us better explanation in the demand for and choice which is less complicated statistical method rather than type II model.

Key Words : tobit, two-step tobit, linear probability model, self-selectivity correction factor, Lambda, statistical efficiency, unbiasedness

I. Introduction

In his book, Introduction to Econometrics, Maddala (1992) has shown a schematic description of several steps involved in an econometric analysis of economic models. He stated that the major purpose of econometrics is to test economic theory, other related and important issues in econometrics such as modeling, estimation method, confirming economic theories, adequacy of suggested model, and usefulness of the model for prediction and policy.

Based on his schema mentioned earlier, the goals of this study are described as follows: first, to construct statistical and econometric models on the basis of economic theory; secondly, to discuss the issue of statistical efficiency and unbiasedness including the sample selection bias correcting problem; thirdly, to compare the estimation methods between one state Tobit of Maximum Likelihood estimation and Heckman's two-step Tobit of Ordinary Least Squares; forthly, to show the adequacy of the model; and, finally, to show the simulation result for prediction and public policy.

Corresponding Author: Sun-Young Park, Department of Venture Technology Management Graduate School of Venture, Hoseo University E-mail: sypark@office.hoseo.ac.kr C.P.: 011-9256-3760 Office: 2055-1401(ext.131)

II. Economic Theory, Econometric Model, and Estimation Method

1. One-stage Tobit Model (The Type I Tobit): $p(y_1<0) \cdot p(y_1)$

According to Amemiya (1984, 1985), a household is assumed to maximize the utility, subject to budget constraint, which is described as follows.

$$\text{Max } U(y, z) \qquad (1)$$
$$\text{Subject to } y + z \leq x$$
$$y \geq 0$$

where y is a household's actual child care expenditures
z is all other expenditures
x is income
$y + z \leq x$ stands for budget constraints, and
$y \geq 0$ stands for boundary constraint.

Also, suppose y^* is the solution of the unconstrained optimization (the utility maximization subject to income constraint only), which is denoted by

$$y_i^* = \beta' xi + \mu_i \qquad (2)$$

Where x_i are the income and other variables and μ_i indicates all the unobservable variables affecting the household's utility.

We can also define y_i^* as desired expenditures or potential expenditures. Thus, we can rewrite equation (2) as

$$y_i = y_i^* \quad \text{if } y_i^* > 0 \qquad (3.1)$$

$$y_i = 0 \quad \text{if } y_i \leq 0 \qquad (3.2)$$

where y_i are observed if a household's potential expenditures are greater than zero ($y_i^* > 0$), y_i^* are unobserved if a household's potential expenditures is less than zero ($y_i^* \leq 0$), x_i are observed variables, and $\mu_i \sim \text{iid } N(o, \sigma^2)$.

Therefore, y_i stands for the actual child care expenditures and x stands for annual earned income, price of child care, financial assets, age of mother, mother's working hours, and other variables.

The likelihood function of the standard Tobit model is given by

$$L = \prod_0 [1 - \Phi(\beta' x_i/\sigma)] \prod_1 \sigma^{-1}\phi[(yi - \beta' x_i/\sigma)] \quad (4)$$

where Φ and ϕ stand for the cumulative distribution function and density function, respectively, of the standard normal distribution; $\prod_0$ stands for the product over those i for which $y_i = 0$ ($y_i^* \leq 0$); and $\prod_1$ means the product over those i for which $yi = 1$ ($yi^* > 0$).

Maddala (1983, 1992) also described the likelihood function for the type I Tobit model which is the same as in description (4), and it takes the following form as:

$$L = \prod_{y^*>0} \frac{1}{\sigma} f[(yi - \beta' x_i/\sigma)] \prod_{y^*\leq 0} F(-\beta' x_i/\sigma) \qquad (5)$$

$$\text{where } f(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

$$F(\beta' x_i/\sigma) = \int_{-\infty}^{\beta' sxi/\sigma} f(t)dt$$

The description of the likelihood function is very important because of the maximum value of the function with respect to β and σ^2, yielding consistent and efficient estimators.

Numerous studies about consumer's purchase of durable goods using the Tobit model use this type of Tobit. Sung, Park, and Hanna (1994) used the type I Tobit model for the determinant of child care.

2. The Type II Tobit: $p(y_1<0)\ p(y_1>0, y_2)$

The previous type 1 Tobit model, p (y_i<0) p (y_i) is another expression for $\prod_0 p(y_i^* <0) \cdot \prod_1 f_i(y_i)$, where f_i is the density function of $N(\beta' x_i, \sigma^2)$. Here, the type II Tobit is defined as p(y_i<0) p (y_i>0, y_2) and follows:

$$y_{1i}^* = \beta_1' x_{1i} + \mu_{1i} \quad (6.1)$$

$$y_{2i}^* = \beta_2' x_{2i} + \mu_{2i} \quad (6.2)$$

where μ_{1i}, μ_{2i} are i, i, d bivariate normal distribution with a mean of zero, variances σ_1^2, σ_2^2 and covariance $\sigma_{12} = \rho\sigma_1\sigma_2$ and,

$$y_{1i} = 1 \text{ if } y_{1i}^* > 0 = 0 \text{ otherwise} \quad (7.1)$$

$$\begin{aligned} y_{2i} &= y_{2i}^* \text{ if } y_{1i} = 1 \\ &= 0 \quad \text{if } y_{1i} = 0 \end{aligned} \quad (7.2)$$

where y_{1i}^* is the households decision on whether to purchase market child care or not, and y_{1i} is observed only when y_{1i}^* >0 indication that y_{1i}^* is the utility difference between purchasing of market child care and non-purchasing market child care, and $y_{1i}$and y_{2i} are the observed variables.

When a household has a child care expense, then y_{1i} takes the value of 1, which means there was a child care expense; otherwise, it takes zero, which means there was no expense. y_{2i}^* is only observed when there is a child care cost; thus, y_{2i} stands for actual child care expenditures only when a household has a child care expense.

Equations (7.1) and (7.2) are probit selection equations, and equations (6.2) and (7.5) are the regression model where the sample selection correction factor, λ, is used as a regressor.

Unlike the type I Tobit, y2i in the type II Tobit can take negative values.

The likelihood function of type II Tobit is given by

$$L = \prod_0 P(y_{1i}^* \leq 0) \prod_1 f(y_{2i}|y_{1i}^*>0)\, p(y_{1i}^*>0). \quad (8)$$

Where $\prod_0$ and $\prod_1$ · stand for the product over these i for which $y_{2i} = 0$ and $y_{2i} \neq 0$, and $\cdot(y_{2i}|y_{1i}^*>0)$ stands for the conditional density of y_{2i}^* given $y_{1i}^*>0$.

The first part of equation (8) is rewritten as

$$\begin{aligned} p(y_{1i}^* \leq 0) &= \textstyle\int_{-\infty}^{0} dy_1 * \int_{-\infty}^{\infty} f(y_1^*, y_2^*) dy_2^* \\ &= \textstyle\int_{-\infty}^{0} f(y_1^*) dy_1^* = \Phi\left(\frac{0 - x_{1i}\beta_1}{\sigma_1}\right) \\ \text{where } & f(y_1^*, y_2^*) = f(y_2^*|y_1^*) f(y_1^*) \\ &\text{and } y_1^* \sim N(x_{1i}\beta_1, \sigma_1^2); \end{aligned} \quad (9.1)$$

and the second part of equation (8) is

$$\begin{aligned} & f(y_{2i}|y_{1i}^*>0)\, p(y_{1i}^*>0) \\ &= \textstyle\int_0^{\infty} f(y_1^*, y_2^*) dy_1^* \\ &= f(y_2) \textstyle\int_0^{\infty} f(y_1^*|y_2^*) dy_1^* \quad (9.2) \\ &= \frac{1}{\sigma^2}\phi\left(\frac{y_2 - x_{2i}\beta_2}{\sigma_2}\right)\phi\left[\frac{x_{1i} + \frac{\rho\sigma_1}{\sigma_2}(y_{2i} - x_{2i}\beta_2)}{\sigma_1\sqrt{1-\rho^2}}\right]. \end{aligned}$$

where

$$f(y_2) = \frac{1}{\sigma^2} \cdot \phi\left(\frac{y_{2i} - x_{2i}\beta_2}{\sigma_2}\right)$$

$$f(y_1^*|y_2) = \frac{1}{\sigma_1\sqrt{1-\rho^2}} \cdot \phi[\frac{y_{1i}-x_{1i}\beta_1 + \frac{\rho\sigma_1}{\sigma_2}(y_{2i}-x_{2i}\beta_2)}{\sigma_1\sqrt{1-\rho^2}}]$$

Therefore by, substituting equations (9.1) and (9.2) into equation (8), the final likelihood function can be produced as follows:

$$L = \prod_0 [\Phi(-\frac{x_{1i}\beta_1}{\sigma_1})] \prod_1 [\frac{1}{\sigma^2}\phi(-\frac{y_{2i}-x_{1i}\beta_2}{\sigma^2}).$$

$$\Phi\{\frac{x_{1i}\beta_1 + \frac{\rho\sigma_1}{\sigma_2}(y_{2i}-x_{2i}\beta_2)}{\sigma_1\sqrt{1-\rho^2}}\}] \quad (10)$$

where $\prod_0$ and $\prod_1$ · stand for the product over these i for which $y_{2i}=0$ and $y_{2i}\neq 0$, and $f(y_{2i}|y_{1i}^*>0)$ stands for the conditional density of y_{2i}^* given $y_{1i}^*>0$.

1) Heckman's Two-step Estimation Method

On the basis of models form (6.1) to (6.2), Heckmans two-step method will be used for the estimation. According to Amemiya (1984, 1985), Maddala (1983) and Cosslett(1994), referring to the models (6.1) to (6.2) again,

$$y_{1i}^* = \beta_1' x_{1i} + \mu_{1i} \quad (6.1)$$

$$y_{2i}^* = \beta_2' x_{2i} + \mu_{2i} \quad (6.2)$$

when we have $E(\mu_{2i}|\mu_{1i})=0$, then the regression method does not have any statistical problem.

$$E[\mu_{2i}|x_{2i}, y_{1i<0}] = E[\mu_{2i}|x_{2i}, \mu_{1i}>\beta_1' x_{1i}]$$

$$\neq 0 \text{ if } \rho \neq 0 \quad (11)$$

where ρ is Cov (μ_{1i}, μ_{2i})

In order to correct the sample selection bias,

$$y_{2i} = \beta_2' x_{2i} + E[\mu_{2i}|\mu_{1i}>-\beta_1' x_{1i}] + v_{2i} \quad (12)$$

where $v_{2i} = \mu_{2i} - E[\mu_{2i}|\mu_{1i}>-\beta_1' x_{1i}]$

Assume μ_{1i} and μ_{2i} are bivariate normal with

$$E[\mu_{2i}|\mu_{1i}>-\beta_1' x_{1i}] = \rho\sigma_2 \frac{\phi_i}{1-\Phi_i} \quad (13)$$

where $\phi_1 = \phi(-\beta_1' x_{1i})$

$\Phi_1 = \Phi(-\beta_1' x_{1i})$

In order to get consistent $\tilde{\beta}_1$, we need to estimate the selection equation by probit in the first step as follows:

$$\tilde{\omega}_i = \frac{\tilde{\phi}_i}{1-\tilde{\Phi}_i}$$

$$\text{where } \tilde{\phi}_1 = \phi(-\tilde{\beta}_1 x_{1i}) \quad (14)$$

$$\tilde{\Phi}_1 = \Phi(-\tilde{\beta}_1 x_{1i})$$

Substitute equation (14) into equation (6.2), then we get

$$y_{2i} = \beta_2' x_{2i} + \gamma\tilde{\omega}_i + \varepsilon_{21} \quad (15)$$

where γ will estimate $\rho\sigma_2$

In the second step, equation (15) will be estimated by using OLS or GLS where the correct term $\gamma\tilde{\omega}_i$ is used for a regressor.

The rewritten equation (15) is as follows:

$$y_{2i} = \beta_2' x_{2i} + \gamma\tilde{\omega}_i + [v_{2i} + \gamma(\omega_i - \tilde{\omega}_i)]. \quad (16)$$

In equation (15), standard error needs to be corrected

$$\tilde{\omega}_i - \omega_i \equiv \frac{\alpha\omega_i}{\alpha\beta_i}(\tilde{\beta}_1 - \beta_1) \quad (17)$$

where the term of $\tilde{\beta}_1 - \beta_1$ will depend on μ_{1i} s.

If γ is statistically significant, it indicates that there was a sample selection bias. Further information on sample selectivity correction is available from Maddala (1983) and Amemiya (1984 and 1985).

III. Research Methodology

1. Data

The most commonly used household spending data in the U. S. A. is the Consumer Expenditure Survey (CES) . The data source in this study was the 1990-1992 CES data. These data have been published by the Bureau of Labor Statistics (BLS).

2. Sample Selection

For the analysis, 791 urban households with at least one child under age six were used. Among the total sample (N=971, sample C), about 37.60 percent of the households (N=365m sample A) had child care expenditures compared to 62.41 percent of households (N=606, sample B) which did not have child care expenditures.

3. Research Model

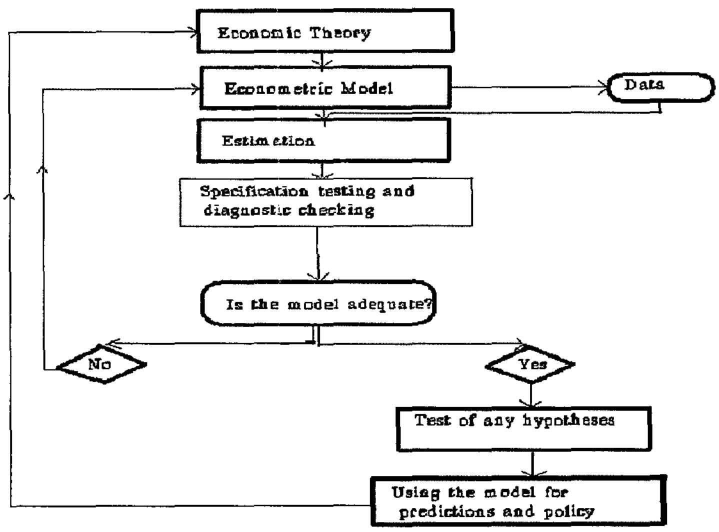
A schematic description of the steps involved in an Econometric Analysis of Economic Models in Maddala, 1992 is in shown Figure 1.

IV. Results

1. One-stage Tobit (M.L.E.)(Table 1)

2. Heckman's Two-step Tobit

1) Probit Selection Model (M.L.E.)(Table 2)



<Figure 1>

<Table 1> Estimates on Demand for Child Care Services, Represented by Potential, Actual Child Care Expenditure and Mean Marginal Effects, Restricted Type I Tobit Model[a]

Variables (x_i)	β	$\partial E(y)/\partial x_i$	$\partial E(y\mid y^{*}>0)/\partial x_i$	p-value
Intercept	-18515.65	-6029.14	-5112.24	.0151
Lnincome	936.27	304.87	258.50	.0001
LnPrice	-1540.62	-501.66	-425.37	.0788
LnFA*	40.90	13.32	11.29	.1490
Hours/week	56.13	18.28	15.50	.0001
Age	1020.43	332.28	281.74	.0504
Agesq	-2615.91	-851.81	-722.27	.0489
Agecub	2130.91	693.88	588.35	.0441
Education (less than high school omitted)				
High	68.69	22.37	18.97	.8682
College	900.40	293.19	248.60	.0766
Morecol	1539.93	501.43	425.17	.0061
Number of Child				
Age <2	441.21	143.67	121.82	.0834
Age 3-5	596.13	194.11	164.59	.0142
Age 6-11	-349.28	-113.73	-96.44	.0661
Age 12-17	-822.84	-267.94	-227.19	.0156
Sigma	3097.10			
Log - L	-3157.00			

LIMDEP was used for calculating three mean marginal effects.

a: Z score= -0.45 , $\Phi(z) = 0.33$, $\phi(z) = 0.36$

*: FA indicates Financial Assets

2) Linear Probability Model with Sample Selectivity Correction Factor (Ordinary Least Square Method)(Table 3)

3. Comparison Between the Type I Tobit and the Type II Tobit

Table 4 shows the comparison between MLE, Type I Tobit, and OLS, Type II Tobit. The results where type I Tobit provides us with more statistical results but none of the explanatory variables in the second step are statistically significant. Such a statistical inefficiency in the second step may be due to the fact that the Lambda, a self-selectivity correction factor, included the same factors in the first selection model. Thus, in the case of child care analysis on demand and by choice, we believe that there is not a significant difference in explanatory variables between the first selection model and the second linear probability model.

Since the Lambda, the self-selectivity correction factor in the Type II Tobit is not statistically

<Table 2> Maximum Likelihood Estimates and Mean Marginal Effects on the probability of Purchasing Market Child Care Services, Probit Selection Model, First Step, Type II Tobit

Variables (Xi)	β	S.E	$\partial p_{(y=1)} / \partial x^1_i$	P-value
Intercept	-7.66500	3.95600	N.A	.0527
Lnincome	.24282	.00705	.0902	.0006
LnPrice	-.70607	.30550	-.2623	.0208
LnFA*	.01025	.00978	.0038	.2942
Hours/week	.00973	.00242	.0036	.0001
Age	.67349	.33720	.2502	.0458
Agesq	-1.8538	.9900	-.6886	.0635
Agecub	1.6253	.96490	.6037	.0921
Education (less than high school omitted)				
High	.08889	.13890	.0330	.5222
College	.31067	.17730	.1154	.0797
Morecol	.46756	.20180	.1737	.0205
Number of Child				
Age <2	-.02800	.09027	-.0104	.7564
Age 3-5	.10304	.08498	.0383	.2253
Age 6-11	-.12635	.06512	-.0469	.0523
Age 12-17	-.25329	.11390	-.0941	.0262
Log-L	-584.48			
Chi-square	116.69			

Marginal effects of independent variables on the probability of purchasing market child care are $\partial p(y=1)/\partial xi = \phi(\beta' x_i)\beta_i$, where $\phi(\cdot)$ is the standard normal probability density function.
* : FA indicates Financial Asset

<Table 3> Ordinary Least Square Estimates On the Demand for Child Care Services, Second Step, Type II Tobit

Variables (x i)	β	S.E	p-value
Intercept	-93856.00	115200.00	.4153
Lnincome	3822.30	3929.00	.3306
LnPrice	-8224.70	11400.00	.4705
LnFA*	170.37	227.20	.4534
Hours/week	183.92	145.60	.2065
Age	5710.40	7977.00	.4741
Age squared	-11505.00	20800.00	.4692
Age cubed	12511.00	17010.00	.4619
Education (less than high school omitted)			
High	1158.20	2722.00	.6705
College	4234.50	5088.00	.4053
Morecol	6284.2	6887.00	.3615
Number of Child			
Age <2	505.45	1535.00	.7419
Age 3-5	1930.70	2023.00	.3398
Age 6-11	-1645.5	2160.00	.4462
Age 12-17	-3607.9	4402.00	.4125
Lambda	18450.00		.3920

*: FA indicates Financial Assets

significant, there is no self-selectivity in the Type II Tobit model, indicating that the Type I Tobit model would give us a better explanation in the demand for and choice of child care. Therefore, in the discussion below, Type I results will be used for prediction and public policy.

4. Specification Testing, Diagnostic Checking, and Model Adequacy

Income and price elasticities of quantity demanded for child care services, using MLE based on Type I Tobit Model are in <Table 5>.

5. Simulation Results for prediction and public policy

Simulation results indicate that child care is a normal good and a necessity for households that already have positive expenses (N= 365), while child care is a luxury good for all households with young children (N=971). The own price elasticity of quantity demanded for child care was -1.65 and -1.26 for both households with total samples (N=971) and positive expense groups (N=365), respectively. This implies that own price effects can be a part of the consumer decision making process.

<Table 4> Estimates of the comparison between Type I Tobit (MLE), and Type II Tobit (OLS)

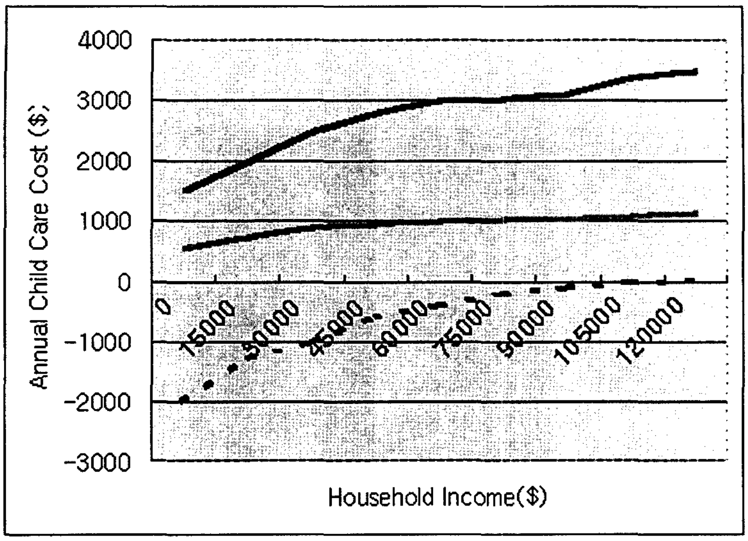
Variables (Xj)	β_{MLE}	$\beta_{HECKMAN}$
Intercept	-18515.00**	-93856.00
Lnincome	936.27***	3822.30
LnPrice	-1540.62*	-8224.70
LnFA*	40.90	170.37
Hours/week	56.13***	183.92
Age	1020.43*	5710.40
Age squared	-2615.91**	-15055.00
Age cubed	2130.91***	12411.00
Education (less than high school omitted)		
High	68.69	1158.20
College	900.40*	4234.50
Morecol	1539.93***	6284.2
Number of Child		
Age <2	441.21*	505.4
Age 3-5	596.13**	1930.70
Age 6-11	-349.28*	-1645.50
Age 12-17	-822.84**	4402.00

* ; P<.1, ** ; P<.05 *** ; P<.01

<Table 5> Income and Price Elasticities

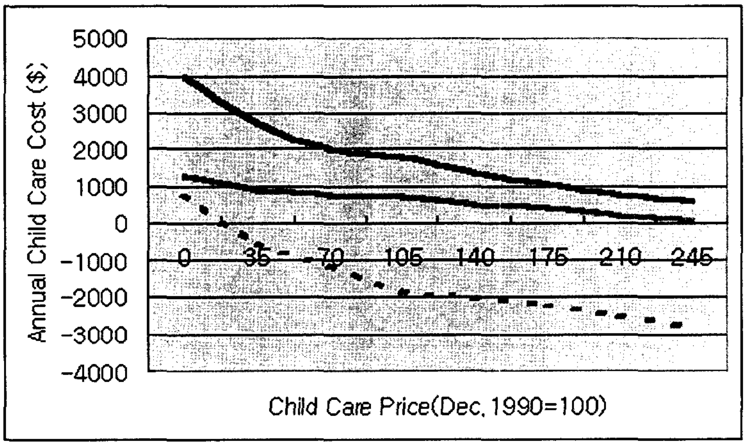
Elasticity	ε_1 (Exp≥0)	ε_2(Exp>0)
Income	0.47	1.24
Price	-1.26	-1.65

In econometric modeling, several problems still remain regarding the estimation method and theories of household demand for and choice of child care services: fertility, decision to work, and choice of purchasing market services simultaneously. In the long-run, we indeed need to attempt another estimation method where the multivariate normal distribution is involved, whereas this paper only tested bivariate normal distribution in econometric and statistical consideration.



<Fig. 2> Simulation Results Based on Tobit Estimates

Effect of Household Income on the Potential, Actual Child Care Cost, holding preferences and relative prices constant, Measured at sample mean level



<Fig. 3> Simulation Results Based on Tobit Estimates

Effect of Child Care Price on the Potential, Actual Child Care Cost, holding incomes and preferences constant, Measured at sample mean level

■ References

Amemiya, T. (1973). Regression Analysis When the Dependent Variable Is Truncated Normal, *Econometrica, 41(6)*, 997-1016.

Amemiya, T. (1981). Qualitative Response Model: A Survey, *Journal of Economic Literature*, 1488.

Amemiya, T. (1984). Tobit Models: A Survey, *Journal of Econometrics, 24*, 3-61.

Amemiya, T. (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

American Chamber of Commerce Researchers Association (1992). *Cost of Living Index.*

Green, W. H. (1990). *Econometric Analysis*, New York, NY: Macmillan Publishing Co.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

Maddala, G. S. (1992). *Introduction to Econometrics,* Second Edition, NY: Macmillan Publishing Company.

Pindyck, R. S., and Rubinfeld, D. L. (1991). *Econometric Models and Economic Forecasts*, 3rd ED. Mcgraw-Hill, Inc.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables, *Econometrica, 26*, 24-36.

Received September 16, 2003
Accepted November 15, 2003