ELSEVIER

Economics Letters 66 (2000) 151–157

economics letters

www.elsevier.com/locate/econbase

Dealing with bottled water expenditures data with zero observations: a semiparametric specification

Seung-Hoon Yoo[a,*], Chang-Young Yang[b]

[a]*Techno-Economics Policy Program, Seoul National University, San 56-1, Shinrim-Dong, Kwanak-Ku, Seoul 151-742, South Korea*

[b]*Division of Economics and International Trade, Hoseo University, 29-1 Sechul-Ri, Baebang-Myun, Asan, Chungnam 336-195, South Korea*

Received 29 March 1999; accepted 19 July 1999

Abstract

This paper analyzes bottled water expenditures data with zero observations by employing parametric and semiparametric models. The overall results of specification tests indicate that the semiparametric model outperforms the parametric model significantly. © 2000 Elsevier Science S.A. All rights reserved.

Keywords: Bottled water expenditures; Two-equation model; Semiparametrics

JEL classification: C34

1. Introduction

Modeling consumer behavior with microeconomic data is complicated by zero observations in the sample. One example is the household expenditures on bottled water, which is rapidly growing in Seoul, South Korea, because of distrust in tap water quality.

To take account of the zero observations, two types of models have been proposed. The first is the censored regression model such as the Tobit model. The specification implies that the stochastic process that determines the binary outcome to participate in consumption also determines the level of the dependent variables. This leaves the burden of differentiating between consumption and non-consumption to the distribution of the single stochastic component and may not be appropriate for modeling zero expenditures in household budgets (Cragg, 1971). Second, a two-equation model, which is very similar to sample selection model (Heckman, 1979) or Type II Tobit model (Amemiya,

*Corresponding author. Tel.: +82-2-880-8701(516); fax: +82-2-880-8389.
E-mail address: yoosh@plaza1.snu.ac.kr (S.-H. Yoo)

0165-1765/00/$ – see front matter
PII: S0165-1765(99)00214-1

1984), has been widely used in recent times. The model incorporates a two-level decision structure, a participation decision and a decision on the amount to spend conditional on deciding to participate, and features two separate stochastic processes that determine the probability and conditional level of expenditure (Melelnberg and Van Soest, 1996).

The econometric analysis of the two-equation model has usually been based on the maximum likelihood (ML) estimation or a two-step estimation of Heckman (1979), assuming the bivariate normality on the distribution of the error terms. However, if the assumption is violated, the estimators are inconsistent. The test, explained in Section 4, can reject the hypothesis of bivariate normality at the 1% level. Thus, the assumption required to use the parametric method is too strong to be satisfied. Although the assumption can be relaxed through the use of different distributional assumptions, it is more appealing to consider a method requiring fewer parametric assumptions.

Recently, a number of semiparametric estimation methods for the two-equation model have been developed. While the econometric theory of these semiparametric estimators has received much attention, empirical applications of the methods remain lacking. This paper, therefore, has two major goals. The first is to investigate household expenditure on bottled water, a commodity in which zero observations may be caused by factors besides true non-consumption. The second is to provide a more careful consideration of the parametric method to show that an alternative semiparametric method can outperform the parametric method.

2. The model

2.1. Parametric model

Our main concern is to estimate a $k_2 \times 1$ parameter vector γ in the two-equation model expressed as:

$$y_{1i} = 1(x_i'\beta + u_{1i} > 0), \tag{1}$$

$$y_{2i} = y_{1i} \cdot (z_i'\gamma + u_{2i}) \tag{2}$$

where $i = 1,2,\ldots,N$ indicates observations, y_{1i} is an indicator variable that denotes whether y_{2i} generated by the regressors z_i is uncensored and that depends on a vector of conditioning variables x_i, β is a $k_1 \times 1$ vector of parameters to be estimated, and u_{1i} and u_{2i} are error terms. The x_i, z_i, β and γ vectors are partitioned into $[1{:}x_{2i}]$, $[1{:}z_{2i}]$, $[\beta_1{:}\beta_2]$ and $[\gamma_1{:}\gamma_2]$, respectively. In the parametric model, the error terms u_{1i} and u_{2i} in the 'participation' Eq. (1) and the consumption 'level' Eq. (2) are assumed to be distributed as bivariate normal.

For observation with $y_{1i} = 1$, the level Eq. (2) takes the form:

$$y_{2i} = z_i'\gamma + \lambda(x_i'\beta) + v_i \tag{3}$$

where $\lambda(x_i'\beta) \equiv E[u_{2i}|u_{1i} > -x_i'\beta]$ and v_i is a new error term. The more frequently employed

parametric method for (3) is Heckman's two-step estimator rather than the ML estimator due to the computational complexity of the latter. At the first step, the parameters β of the participation Eq. (1) are estimated by the probit ML estimation, using the representation:

$$\Pr(y_{1i} = 1|x_i) = E[y_{1i}|x_i] = F(x_i'\beta) \tag{4}$$

where $F(\cdot)$ is the cdf of $-u_{1i}$, specified as standard normal. Second, given the estimator $\hat{\beta}$ from the first step, the estimator $\hat{\gamma}$ is obtained by least-squares (LS) estimation after replacing $\lambda(x_i'\beta)$ with $\phi(x_i'\hat{\beta})/\Phi(x_i'\hat{\beta})$ where ϕ and Φ are standard normal pdf and cdf, respectively. As discussed in Section 1, if the joint distribution of the error terms is misspecified, the second-step estimator will be inconsistent.

2.2. Semiparametric model

As an alternative to the parametric method, Newey et al. (1990) suggested a semiparametric method whose strategy is the same as that of Heckman's two-step method. Taking $E(\cdot|x_i'\beta)$ on (3) and then subtracting this from (3) produces:

$$y_{2i} - E[y_{2i}|x_i'\beta] = (z_i - E[z_i|x_i'\beta])'\gamma_2 + v_i \tag{5}$$

which does not have $\lambda(x_i'\beta)$ any more. The intercept term (γ_1) disappears due to the mean subtraction. The first step in the semiparametric method is to obtain a consistent estimator of β. Several $\sqrt{N}$-consistent and asymptotically normal estimators of β have been developed. Of the estimators, the estimator of Klein and Spady (1993), employed in this study, achieves the semiparametric efficiency bound of Cosslett (1987) for the binary choice model under regularity conditions and with the scale/sign normalization a parameter to one.

The estimator is called a quasi-ML estimator (QMLE), which selects $\tilde{\beta}$ to maximize

$$\frac{1}{N}\sum_{i=1}^{N}\left[y_{1i}\ln\hat{F}(x_i'\beta) + (1 - y_{1i})\ln\{1 - \hat{F}(x_i'\beta)\}\right], \tag{6}$$

where $\hat{F}(x_i'\beta)$ is a nonparametric estimate of $F(x_i'\beta)$. As in Klein and Spady (1993), $\hat{F}(x_i'\beta)$ is calculated from nonparametric kernel estimates, taking the form:

$$\hat{F}(x_i'\beta) = \frac{\sum_{j=1}^{N} K\left(\frac{(x_i - x_j)'\beta}{h}\right) y_{1j}}{\sum_{j=1}^{N} K\left(\frac{(x_i - x_j)'\beta}{h}\right)}, \tag{7}$$

where $K(\cdot)$ is a kernel function and h is a bandwidth parameter.

Given the QMLE $\tilde{\beta}_2$, the nonparametric estimators $\hat{E}[y_{2i}|x_{2i}'\tilde{\beta}_2]$ and $\hat{E}[z_i|x_{2i}'\tilde{\beta}_2]$ for $E[y_{2i}|x_i'\beta]$ and $E[z_i|x_i'\beta]$ are obtained using kernel estimators of the same form described in (7), replacing y_{1j} by y_{2j}

and z_{2j}.[1] The second step can define a new dependent variable $y_{2i} - \hat{E}[y_{2i}|x'_{2i}\tilde{\beta}_2]$ and regressors $z_{2i} - \hat{E}[z_{2i}|x'_{2i}\tilde{\beta}_2]$ in (5) and then apply LS to the new model.

3. Empirical results

This study covers household at-home consumption of bottled water, using the data collected from a survey in Seoul in 1997. The sample consists of 500 housekeeping households. The definitions and sample statistics of the variables used in the study are presented in Table 1. The sample is censored, with 421 households (84.2%) reporting non-consumption of bottled water.

For the semiparametric estimates, the kernel function $K(\cdot)$ was taken as a standard normal density function; the bandwidth parameter h was determined by generalized cross-validation criterion. The data-driven procedure eliminates the arbitrariness in choosing the bandwidth parameter and has certain optimality properties under suitable conditions.

Table 2 shows the coefficients of our basic equations estimated by the parametric and semiparametric methods. The coefficient of STAIN is 1.0 in both the parametric and the semiparametric binary choice models by scale/sign normalization. The standard errors of the estimates in level equations were computed using the heteroscedasticity-consistent and the first step's approximation error-

Table 1
Description of variables in model

Variable	Definition	Mean	Standard deviation
CONSUME	Monthly expenditure for bottled water consumption with zero observations (unit: 1000 won[a])	3.572	9.257
QUALITY	Opinion concerning current tap water quality (0=bad; 1=moderate; 2=good)	0.780	0.648
STAIN	Dummy for experience of bad tap water quality such as rust stain and odor (0=no; 1=yes)	0.348	0.477
PWATER	Dummy for using purifier (0=no; 1=yes)	0.154	0.361
CHILD	Dummy for having a child under 13 (0=no; 1=yes)	0.584	0.493
FAMILY	Log of number of family	1.333	0.289
RESIDE	Log of number of years respondent has been a resident of Seoul	2.822	0.860
INCOME	Log of monthly household total income after tax deduction (unit: 10 000 won)	5.276	0.478

[a] At the time of the survey, $US 1 is approximately equal to 800 Korean won.

[1] The usual intercept term (β_1) in (6) is not estimable because it disappears in (7).

Table 2
Estimation results[a]

Variables	Coefficients			
	Parametric		Semiparametric	
	Participation	Level	Participation	Level
CONSTANT	−6.2264	−172.6376		
	(0.9466)	(94.4277)		
QUALITY	−0.1348	−9.0602	−0.1115	−0.6304
	(0.1185)	(5.5146)	(0.0087)	(1.9256)
PWATER	−0.5620	−23.2870	−0.5742	−24.5690
	(0.2468)	(12.2009)	(0.0139)	(7.5843)
CHILD	0.3414	12.0055	0.2966	15.4358
	(0.1629)	(5.8951)	(0.0098)	(5.3634)
FAMILY	−0.5707		−0.5229	
	(0.2840)		(0.0176)	
RESIDE	0.2566	6.9073	0.2398	9.0856
	(0.0969)	(3.9365)	(0.0050)	(3.6820)
INCOME	0.8859	21.7257	0.8422	6.0832
	(0.1710)	(11.7000)	(0.0114)	(4.1104)
STAIN	1.0000	9.8256	1.0000	4.2284
		(5.7607)		(2.5309)
λ		36.5303		
		(17.4248)		
Log-likelihood[b]				
R^2	−198.46		−154.54	
		0.230		0.860

[a] Standard errors are reported in parentheses below the estimates.
[b] Log-likelihood for the parametric model and quasi-log-likelihood for the semiparametric model.

consistent covariance formulas suggested by Ryu (1996). The semiparametric model explains substantially more variations in the dependent variable than the parametric model, $R^2 = 0.860$ vs. 0.230.[2]

4. Specification tests

There are a number of tests that compare parametric with semiparametric models (see Lee, 1996). Here, two tests are employed. First, we use a specification test of Whang and Andrews (1993) that has power against the violation of bivariate normality assumption in the parametric model. The test statistic is asymptotically distributed as chi-squared. The computed test statistic is 31.33, which is large enough to reject the null hypothesis of the bivariate normality at the 1% level, given that $\chi^2_{0.01}(13) = 27.69$.

[2]The R^2 of the semiparametric model is calculated as $\Sigma_{i=1}^N \hat{y}_{2i}^2 / \Sigma_{i=1}^N$ where $\hat{y}_{2i} = \hat{E}[y_{2i} | x_{2i}'\tilde{\beta}_2] + (z_{2i} - \hat{E}[z_{2i} | x_{2i}'\tilde{\beta}_2])'\tilde{\gamma}_2$. It might well be a useful descriptive device for comparing models since the approach still produces a value between zero and one when the constant term is not contained.

Second, we conduct Hausman (1978) type specification test given in Robinson (1988):

$$H_N = (\hat{\gamma}_2 - \tilde{\gamma}_2)' \Omega^{-1} (\hat{\gamma}_2 - \tilde{\gamma}_2) \sim \chi^2(q_H), \tag{8}$$

where Ω is the variance–covariance matrix of $(\hat{\gamma}_2 - \tilde{\gamma}_2)$ and $q_H = k_2 - 1$. The null hypothesis for H_N is that the parametric model estimates would be consistent. The alternative hypothesis is that the model is semiparametric. The calculated statistic is 39.21. Given that $\chi^2_{0.01}(6) = 16.81$, the null hypothesis that the correct specification is the parametric model is rejected at the 1% level. The test result is rather dramatic evidence of the impact of changing estimator in a situation that the bivariate normality assumption on the distribution of the error terms is not satisfied.

5. Conclusions

The parametric estimation technique for a two-equation model such as Heckman's two-step method is vulnerable to non-normal error structure. The semiparametric method is robust under that stress. In the application reported here, the semiparametric model outperformed the parametric model significantly, reducing the implicit restrictions involved in the parametric model. The empirical estimation of the semiparametric two-equation model has not been popular since the estimator requires cumbersome nonparametric regressions. With advances in computer technology, it can be easily calculated these days. Therefore, it is a practical as well as a theoretically promising way of dealing with nonexperimental microeconomic data with zero observations.

Acknowledgements

The authors wish to thank Y.J. Whang for his helpful comments on earlier versions of this paper. All errors remain the authors' responsibility.

References

Amemiya, T., 1984. Tobit models: a survey. Journal of Econometrics 24, 3–61.

Cosslett, S.R., 1987. Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. Econometrica 55, 559–585.

Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica 46, 1251–1271.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Klein, R.L., Spady, R.H., 1993. An efficient semiparametric estimator for discrete choice models. Econometrica 61, 387–422.

Lee, M.J., 1996. In: Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models, Springer, New York, NY.

Melelnberg, B., Van Soest, A., 1996. Parametric and semiparametric modeling of vacation expenditures. Journal of Applied Econometrics 11, 59–76.

Newey, W.K., Powell, J.L., Walker, J.R., 1990. Semiparametric estimation of selection models: some empirical results. American Economic Association Papers and Proceedings 80, 324–328.

Robinson, P.M., 1988. Root-*N*-consistent semiparametric regression. Econometrica 56, 931–954.

Ryu, K., 1996. Consistent, positive definite covariance matrix estimation of Heckman's two-step estimators. Journal of Economic Theory and Econometrics 2, 65–76.

Whang, Y.J., Andrews, D.W.K., 1993. Tests of specification for parametric and semiparametric models. Journal of Econometrics 57, 277–318.