A consistent semiparametric estimation of the consumer surplus distribution

Andrew Foster, Jinyong Hahn*

Department of Economics, Brown University, Box B, Providence, RI 02912, USA

Received 3 December 1999; accepted 25 April 2000

Abstract

In this paper, we examine the consequences of demand-function heterogeneity for the estimation of the consumer surplus. In particular, we show that, given a linear demand function with random coefficients, one can consistently estimate the consumer surplus distribution without making parametric assumptions about the coefficient distribution. The approach is illustrated using data on gasoline consumption. © 2000 Elsevier Science S.A. All rights reserved.

Keywords: Consumer surplus; Semiparametric estimation

JEL classification: C2 (Economic methods: single equation models); DO (Microeconomics: general)

1. Introduction

In recent years, there has been increased recognition that, although modelling consumer heterogeneity in terms of intercept heterogeneity will suffice for many purposes, there are other purposes for which this approach is inadequate. For example, the presence of heterogeneous consumers necessitates a different way of thinking about consumer surplus. Suppose that each consumer's (Marshallian) demand function is known up to some finite dimensional parameter, say $f(p, y, \theta)$: with price vector p and the income y , the consumer will spend $f(p, y, \theta)$. If the consumer heterogeneity is solely summarized by θ , we can then view θ as a random vector, and consumer i 's demand function given θ can be written as $f(\cdot, \cdot, \theta_i)$. Let $S(\theta_i, p^0, p^1)$ denote the consumer surplus change corresponding to the price change from p^0 to p^1 for the consumer with $\theta_i = \theta$. The price change from p^0 to p^1 would imply a consumer surplus change of $S(\theta_i, p^0, p^1)$ for the i th consumer. A Benthamite utilitarian social

*Corresponding author. Tel.: +1-401-863-3883; fax: +1-401-863-1970.

E-mail address: jinyong-hahn@brown.edu (J. Hahn).

planner, whose main concern is the total welfare, would want to know the average consumer surplus $E[S(\theta_i, p^0, p^1)]$. If (p, y) and θ are independent of each other in the population, and if an econometrician assumes consumer homogeneity, then she would estimate $\mathbf{f}(p, y) \equiv E[f(p_i, y_i, \theta_i) | (p_i, y_i) = (p, y)]$, and then report the corresponding consumer surplus change. For example, if the demand function is linear in θ_i , then the econometrician might estimate $E[\theta_i]$, and report $S(E[\theta_i], p^0, p^1)$ as the consumer surplus change. This could lead to inappropriate inferences because in general $E[S(\theta_i, p^0, p^1)] \neq S(E[\theta_i], p^0, p^1)$. The purpose of this paper is to address this problem using nonparametric methods. Specifically, we employ the approaches of Beran and Hall (1992) and Beran and Millar (1994) to analyze consumer surplus in a random coefficients (log) linear demand system.

We consider semiparametric estimation of the consumer surplus distribution and the corresponding average consumer surplus assuming that the θ_i distribution is nonparametrically specified. For simplicity, we only consider the two-good linear demand system:

$$q = f(p, y, \theta_i) \equiv \alpha_i + \beta_i \cdot p + \gamma_i \cdot y \quad (1)$$

where, q , p and y denote the logarithms of quantity demanded, price and income, respectively. We identify and consistently estimate the θ_i distribution utilizing the methodologies of Beran and Hall (1992) and Beran and Millar (1994). With θ distribution identified, it trivially follows that $S(\theta_i, p^0, p^1)$ is identified as long as $S(\theta_i, p^0, p^1)$ can be numerically evaluated. Determining $S(\theta_i, p^0, p^1)$ in general involves solving a partial differential equation, but a closed form expression for our specification is available from Hausman (1981).

2. Identification and estimation

In this section, the identification and consistent estimation of the consumer surplus distribution is discussed for the two-good linear demand system (1). The estimation strategy is taken from Beran and Millar (1994). We first introduce some notation. Let F and G denote the distributions of $\theta_i = (\alpha_i, \beta_i, \gamma_i)$ and (p_i, y_i) , respectively. Let $\mathcal{L}(F, G)$ denote the joint distribution of (q_i, p_i, y_i) under F and G . Define F_n , G_n , and $\mathcal{L}(F_n, G_n)$ to be some sequence of distributions of θ_i , (p_i, y_i) , and (q_i, p_i, y_i) . Finally, let d denote any metric that metrizes weak convergence, e.g. the L_2 -norm on characteristic functions. The minimum distance estimator is defined as follows. Let $\hat{\mathcal{L}}_n$ and $\hat{G}_n$ denote the empirical distributions of (q_i, p_i, y_i) and (p_i, y_i) , respectively. Let m_n denote any sequence of positive integers such that $\lim_{n \rightarrow \infty} m_n = \infty$. Let $M(m_n)$ denote the set of all multinomial distributions in some compact set $\mathbf{C}$ with mass at each point being m_n^{-1} . The minimum distance estimator defined by

$$\tilde{F}_n \equiv \operatorname{argmin}_{F_n \in C(m_n)} d(\mathcal{L}(\hat{F}_n, \hat{G}_n), \hat{\mathcal{L}}_n) \quad (2)$$

can be shown to be consistent by Beran and Millar (1994, Proposition 2.4). Note that the number of support points of this multinomial distribution can be interpreted as the number of consumer ‘types’.

A multinomial distribution with three support points is interpreted as indicating that there are three different types of consumers, for example. Thus, the determination of m_n in the finite sample may be guided by both economic and statistical intuition.

The estimation of the minimum distance estimator requires a choice of $d(\cdot, \cdot)$. For this purpose, we can use the L_2 -norm on the characteristic functions, for example. Then we have $d(P_1, P_2) = (\int |\phi_1(t) - \phi_2(t)|^2 dQ(t))^{1/2}$, where $\phi_1(t)$ and $\phi_2(t)$ are characteristic functions of P_1 and P_2 , and Q is some probability distribution with support equal to the whole Euclidean space. It is often impossible to obtain an analytic expression of $d(\cdot, \cdot)$ when Q has the whole Euclidean space as its support. But we may instead use $d_N(P_1, P_2) = (\int |\phi_1(t) - \phi_2(t)|^2 dQ_N(t))^{1/2}$, where Q_N is the empirical distribution of a random sample of size N from Q . As long as $N \rightarrow \infty$ as $n \rightarrow \infty$, the corresponding simulated minimum distance estimator would still be consistent.

3. Empirical application

To estimate the average consumer surplus generated by changes in gasoline prices, we employ the data used by Hausman and Newey (1995). The data set is from the US Department of Energy, and contains monthly gasoline consumptions q_i , the weighted averages of the gasoline price over a month p_i , incomes y_i , and other personal characteristics x_i of 18 109 observations. These personal characteristics x_i consist of 20 time and region indicator variables. For a more complete description of the data set, see Hausman and Newey (1995, p. 1459). We use the log linear individual demand specification

$$\log q_i = \alpha_i + \beta_i \log p_i + \gamma_i \log y_i \quad (3)$$

We assume that the personal characteristics x_i can be used to predict $\theta_i = (\alpha_i, \beta_i, \gamma_i)$, but the prediction is not perfect: in particular, we assume that $\alpha_i = x_i' \alpha + a_i$, $\beta_i = x_i' \beta + b_i$, $\gamma_i = x_i' \gamma + c_i$, where the vector (a_i, b_i, c_i) has mean equal to zero and is independent of x_i . The distribution of (a_i, b_i, c_i) is assumed to be unknown. If (α, β, γ) is known, we can rewrite the individual demand function as

$$\log q_i^* \equiv \log q_i - x_i' \alpha - (x_i' \beta) \cdot \log p_i - (x_i' \gamma) \cdot \log y_i = a_i + b_i \log p_i + c_i \log y_i \quad (4)$$

and consistent estimation of the (a_i, b_i, c_i) distribution can be achieved by the strategies discussed in the previous sections. Because (α, β, γ) is consistently estimated by the OLS regression of $\log q_i$ on $(x_i, x_i \log p_i, x_i \log y_i)$, the (false) assumption that (α, β, γ) is known and equal to its OLS estimator does not create any problem in the consistent estimation of the (a_i, b_i, c_i) distribution. For the evaluation of the simulated metric $d_N(\cdot, \cdot)$, we need to choose Q and the simulated random sample size N : we chose $Q = N(0, I_3)$, and $N = 1000$. We also need to choose m , the number of support points of $\tilde{F}_n$: we experimented with $m = 4, \dots, 10$.

Our findings are summarized in Tables 1–7. In general, we find that there is a substantial difference between the average equivalent variations computed under the two assumptions. For the price change from \$1.00 to \$1.30, we find that most of the difference is accounted for by the difference in the tax revenue calculation, not to the deadweight loss. One possible explanation is that the difference in deadweight loss is minimal because the deadweight loss is of second order whereas the tax revenue is

Table 1
Estimates of Consumer Surplus with $m=4$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	317.46	513.85	290.34	453.84
	σ	134.57	238.84	82.88	129.85
Tax revenue	μ	294.07	505.82	262.12	388.42
	σ	160.89	360.40	74.52	110.42
Deadweight loss	μ	23.39	8.03	28.22	65.42
	σ	91.39	199.75	8.36	19.43

Table 2
Estimates of Consumer Surplus with $m=5$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	318.56	517.43	288.99	452.73
	σ	140.96	252.30	82.89	130.15
Tax revenue	μ	297.44	513.10	261.94	389.85
	σ	167.76	383.48	74.83	111.37
Deadweight loss	μ	21.09	4.33	27.05	62.87
	σ	89.51	209.88	8.06	18.78

of first order in the price change. For the price change from \$1.00 to \$1.50, though, we observe a different pattern. The difference between these two specifications seems to arise from both sources. The average deadweight loss under unobserved heterogeneity is even lower than the corresponding figure for the price change to \$1.30! Especially troubling is the fact that, for some types of consumers, the deadweight losses are estimated to be negative!¹ We do not have a good explanation, although one

Table 3
Estimates of Consumer Surplus with $m=6$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	320.51	520.22	287.88	451.66
	σ	149.71	261.14	82.84	130.27
Tax revenue	μ	299.03	515.06	261.62	390.52
	σ	169.48	397.80	74.98	111.93
Deadweight loss	μ	21.47	5.16	26.26	61.14
	σ	91.22	227.92	7.86	18.34

¹We verified that the estimated Hicksian demand elasticity was of the wrong sign for those types of consumer.

Table 4
Estimates of Consumer Surplus with $m=7$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	334.00	528.39	293.18	451.10
	σ	186.32	307.77	83.52	128.80
Tax revenue	μ	302.77	478.56	257.20	369.34
	σ	191.12	334.13	72.97	104.79
Deadweight loss	μ	31.23	49.83	35.98	81.76
	σ	63.67	120.97	10.55	24.01

Table 5
Estimates of Consumer Surplus with $m=8$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	326.41	527.52	287.08	450.73
	σ	168.02	287.14	82.48	129.80
Tax revenue	μ	303.57	514.30	261.22	390.48
	σ	180.88	416.68	74.75	111.74
Deadweight loss	μ	22.84	13.22	25.85	60.25
	σ	88.66	234.40	7.73	18.06

possible reason may be that the standard deviation of the deadweight loss is so big that a reliable average is hard to obtain even with a large sample. This is supported by our estimates of the standard deviation of the deadweight loss. A related possibility is that our estimator does not have a fast rate of convergence: even though we have established the consistency of our procedure, we do not expect the estimators to be $\sqrt{n}$ -consistent. Last, it may simply be the case that the log linear demand system is not a good specification for individual demand.

Table 6
Estimates of Consumer Surplus with $m=9$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	329.59	531.72	286.88	451.55
	σ	177.04	301.83	82.62	130.05
Tax revenue	μ	306.24	514.85	261.17	390.62
	σ	188.17	432.50	74.92	112.04
Deadweight loss	μ	23.35	16.87	25.71	59.93
	σ	86.68	240.89	7.70	18.01

Table 7

Estimates of Consumer Surplus with $m=10$

		Unobserved heterogeneity		Observed heterogeneity	
		\$1.0 – \$1.3	\$1.0 – \$1.5	\$1.0 – \$1.3	\$1.0 – \$1.5
Equivalent variation	μ	332.69	534.82	287.16	450.27
	σ	184.14	313.78	82.69	129.95
Tax revenue	μ	308.27	510.98	260.69	388.68
	σ	194.59	434.93	74.77	111.47
Deadweight loss	μ	24.41	23.86	26.47	61.59
	σ	83.26	231.42	7.92	18.48

It is instructive to compare our estimates with Hausman and Newey's (1995). Their equivalent variation estimates are between \$278.95 and \$302.75 for the price change to \$1.30, and \$438.01 and \$475.91 for the price change to \$1.50.² Their deadweight loss estimates are between \$29.19 and \$38.68, and between \$45.80 and \$51.05, respectively. These numbers roughly correspond to our estimates calculated under the assumption that all heterogeneity is observed. On the other hand, our equivalent variation estimates (computed under the assumption that some heterogeneity is unobserved) are between \$317.46 and \$332.69 for the price change to \$1.30, and \$513.85 and \$534.83 for the price change to \$1.50. This difference seems to suggest that the unobserved heterogeneity may be more important than the demand function specification. Our deadweight loss estimates for the price change to \$1.30 are roughly comparable to their numbers, whereas the ones for the price change to \$1.50 are not. Again, we do not very well understand why the numbers are so much different for the latter price change. Focusing on the former price change, though, we observe that most of the difference between our numbers and Hausman and Newey's (1995) can be attributed to the tax revenue calculation.

Acknowledgements

We have benefitted from helpful comments by Joshua Angrist, Valentina Corradi, Christian Gourieroux, and the seminar participants of the University of Pennsylvania, Harvard/MIT, and CentER. We are grateful to Jerry Hausman and Whitney Newey for kindly sharing the data.

References

- Beran, R., Hall, P., 1992. Estimating coefficient distributions in random coefficient regressions. *Annals of Statistics* 20, 1970–1984.

²We cannot make a direct comparison of their numbers with ours. They have computed the equivalent variation at a median income level, whereas we computed the average of the equivalent variation over different types of consumers and over different incomes levels.

- Beran, R., Millar, P., 1994. Minimum distance estimation in random coefficient regression models. *Annals of Statistics* 22, 1976–1992.
- Hausman, J., 1981. Exact consumer's surplus and deadweight loss. *American Economic Review* 71, 662–676.
- Hausman, J., Newey, W., 1995. Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica* 63, 1445–1476.