JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ., **13**, 435–461 (1998)

SEMIPARAMETRIC ESTIMATION AND CONSUMER DEMAND

RICHARD BLUNDELL[a], ALAN DUNCAN[b]* AND KRISHNA PENDAKUR[c]

[a]*University College London and Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK*
[b]*Department of Economics and Related Studies, University of York, Heslington, York YO1 5DD, UK*
[c]*Department of Economics, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6*

SUMMARY

This paper considers the implementation of semiparametric methods in the empirical analysis of consumer demand. The application is to the estimation of the Engel curve relationship and uses the British Family Expenditure Survey. Household composition is modelled using an extended partially linear framework. This is shown to provide a useful method for pooling non-parametric Engel curves across households of different demographic composition. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Consumer demand presents an important area for the application of semiparametric methods. In the analysis of the cross-section behaviour of consumers, non-parametric analysis of the Engel curve relationship is now common place (see Bierens and Pott-Buter, 1990; Härdle and Jerison, 1991; Banks, Blundell, and Lewbel, 1997, for example). The contribution of the present paper is to extend this work in two directions. First, we consider the semiparametric specification of demographic composition to the non-parametric Engel curve relationship. Second, we test some popular parametric specifications for Engel curves against these semiparametric alternatives.

As a baseline specification we work with the Working–Leser or Piglog specification in which budget shares are linear in the log of total expenditure (see Muellbauer, 1976; Deaton and Muellbauer, 1980a). This form for the Engel curve relationship also underlies the popular Almost Ideal and Translog demand models of Deaton and Muellbauer (1980a) and Jorgenson, Lau, and Stoker (1980). Moreover, it provides a useful parametric null hypothesis for the non-parametric alternative. Recent attention has focused on Engel curves which have more variety of curvature than is permitted by the Piglog. This reflects growing evidence from a series of empirical studies that suggest quadratic logarithmic income terms are required for certain budget share equations (see, for example, Atkinson, Gomulka, and Stern, 1990; Hausman *et al.*, 1991; Hausman, Newey, and Powell, 1995; Härdle and Jerison, 1991; Lewbel, 1991; Blundell and Duncan, 1998; and Blundell, Pashardes, and Weber, 1993). Consequently we use both the Piglog and quadratic logarithmic specifications as null parametric specifications for designing tests against a non-parametric alternative.

There are many reasons why it is important to recover an accurate specification of the Engel curve relationship. First, accurate specification is important in modelling consumer responses to,

* Correspondence to: Dr Alan Duncan, Department of Economics and Related Studies, University of York, Heslington, York YO1 5DD, UK. E-mail: asd1@york.ac.uk

Contract grant sponsor: ESRC Centre for the Micro-Economic Analysis of Fiscal Policy at IFS.

CCC 0883–7252/98/050435–27$17·50

Received 15 October 1997
Revised 15 May 1998

and the welfare impact of, policy reforms. Second, for estimating the impact of demographic change and equivalence scales, the shape of Engel curves is critical. As a final motivation for Engel curve analysis we can point to the importance of measuring expansion paths. That is the effect of changes in overall budget on the relative demand for commodities. This plays a central role in the modern analysis of revealed preference on micro-data (see, for example, Blundell, Browning, and Crawford, 1997).

Restrictions from consumer theory are not innocuous both on the form of the Engel curve relationship and on the way in which observable heterogeneity (demographics in our case) can enter. In a non-linear Engel curve, if demographics are to enter in a partially linear semi-parametric specification, then they must in general also scale total expenditure on the right-hand side of the budget share regression. This is equivalent to translating the log of total expenditure that appears as the regressor in the non-parametric generalizations of the Working–Leser specification. Therefore, if we wish to interpret the demographic composition variables as 'taste shifters' in a preference-consistent way, the popular partially linear specification of Robinson (1988) has to be generalized.

The simple generalization, achieved by scaling total expenditure, corresponds to the 'base-independent' (or 'equivalence scale exactness') method of introducing demographics in demand analysis (see Blackorby and Donaldson, 1994, for example). Interestingly this partially linear 'translation' has the same form considered in the pooling of 'shape invariant' non-parametric regression curves of Härdle and Marron (1990) and Pinkse and Robinson (1995), recently explored in the context of equivalence scales by Pendakur (1998).

The shape of Engel curves and consistency with consumer theory is a topic investigated in great detail by Gorman (1981). In general there is no restriction on the shape of Engel curves provided relative prices and demographics are allowed to enter in a completely flexible way. However, if we have in mind to restrict the way prices (or demographics) come in through some parametric specification, then the form of the Engel curve is also restricted through the homogeneity and Slutsky symmetry conditions which tie the expenditure shares and the price and expenditure derivatives closely together. For example, Banks, Blundell, and Lewbel (1997), using the results of Gorman (1981), show that if we consider budget share Engel curves that are additive in a constant, a linear logarithmic term and some function of total expenditure then the demand system is restricted to the quadratic logarithmic family. In general the Working–Leser specification which has shares linear in log total expenditure has been found to provide a close approximation for some goods. In this paper we show that, if demographic composition enters the budget share Engel curves in an additive way, as in the partially linear framework, then consistency with homogeneity and Slutsky symmetry imposes strong restrictions. In particular, if any one good has a Working–Leser Engel curve then all goods are restricted to be Working–Leser. This is a strong restriction that is relaxed in our extended partially linear model.

In the empirical analysis of Engel curves a further important issue is the endogeneity of total expenditure. Since total expenditure may well be jointly determined with expenditure shares it is likely to be endogenous. If total expenditure is endogenous for individual commodity demands, then the conditional mean estimated by non-parametric regression will not identify the 'structural' Engel curve relationship. That is, the 'statistical' Engel curve will not recover the shape necessary for the analysis of consumer preferences, equivalences scales or expansion paths. However, given the two-stage budgeting of choices under separability, the system of budget shares and total expenditure forms a triangular or recursive system and is open to fairly

simple estimation techniques. To account for endogeneity we adapt the Holly and Sargan (1982) augmented regression approach to semiparametric regression context. We also consider the Newey, Powell, and Vella (1995) extension to additive recursive structures.

To compare these semiparametric specifications with the Working–Leser and quadratic logarithmic parametric specifications we implement a recently developed specification test by Aït-Sahalia, Bickel, and Stoker (1994) for this hypothesis (see also Härdle and Mammen, 1993; Ellison and Ellison, 1992; and Zheng, 1996). This analysis shows a strong rejection of the Working–Leser or Piglog form for some budget shares, even after adjusting for demographic differences and endogeneity. However, the quadratic logarithmic model is not rejected. We also test the shape invariance of budget shares across demographic types. For this we implement the smooth conditional moment bootstrap method of Gozalo (1997).

The structure of the paper is as follows. Section 2 takes a look at the shape of Engel curves for a subsample of households in the British Family Expenditure Survey. Section 3 goes on to consider the specification for demographic composition in budget share Engel curves and investigates the restrictions that result from the homogeneity and Slutsky conditions. We consider the shape-invariant extension to the partially linear semiparametric Engel curve model that relaxes the restrictions placed on preferences by the additive structure of demographic and income terms in the partially linear model. Section 4 considers suitable corrections for endogeneity of total expenditure and then applies these ideas to the Engel curve analysis and reports results for the test of Piglog and quadratic logarithmic specifications against semiparametric alternatives. Section 5 concludes.

2. THE SHAPE OF ENGEL CURVES

2.1 The Working–Leser Specification

For most of our analysis we will be concerned with assessing and generalizing the simple relationship between budget shares and total expenditure. These models have the structure

$$w_{ij} = g_j(\ln\, x_i) + \varepsilon_{ij} \tag{1}$$

where w_{ij} is the budget share of the jth good for individual i, $\ln\, x_i$ is the log of total expenditure and the unobservable ε_{ij} is assumed to satisfy $E(\varepsilon_{ij} \mid x_i) = 0$. Choosing to model budget shares in terms of the log of total outlay follows from the original statistical analysis of budget shares by Leser (1963) and Working (1943). It is also motivated by the popular Almost Ideal and Translog demand models of Deaton and Muellbauer (1980a) and Jorgenson, Lau, and Stoker (1980) which also have the 'Piglog' specification in which shares are linear in log total outlay. This form of the Engel curve is commonly refered to as the Working–Leser specification.

2.2 Data Used in this Study

In our application we consider six broad categories of goods; food, domestic fuel, clothing, alcohol, transport, and other goods. We draw data from the 1980–1982 British Family Expenditure Surveys (FES) and, for the purposes of our study, we select only households with one or two children. Total expenditure and income are measured in £ per week. In order to preserve a degree of demographic homogeneity in all aspects other than the number of children in the household, we select from the FES a subset of married or cohabiting couples with an employed head of

Table I. Descriptive statistics for budget share data

	Couple with one child		Couple with two children	
Variable	Means	Std deviations	Means	Std deviations
Food share	0·343	0·109	0·365	0·101
Fuel share	0·093	0·053	0·090	0·051
Clothing share	0·106	0·098	0·108	0·093
Alcohol share	0·067	0·069	0·056	0·059
Transport share	0·138	0·109	0·129	0·102
Other good share	0·253	0·104	0·252	0·103
Total expenditure	94·74	45·84	101·22	41·12
Total net income	134·22	70·45	137·46	54·28
Log total expenditure	4·46	0·41	4·55	0·37
Log net income	4·81	0·40	4·86	0·36
Age of household head	35·70	9·40	35·83	6·52
Sample size	594		925	

household living in Greater London or south-east England. All those who are self-employed, retired or in full-time education are excluded from the sample. This leaves us with 1519 observations, including 925 couples with two children. Table I gives brief descriptive statistics for the main variables used in the empirical analysis.

2.3 Some Picture of the Expenditure Share–Log Total Expenditure Relationship

In Figures 1 to 6 we present kernel regressions of the Engel curves for the six budget shares in our FES sample. Each figure presents unrestricted non-parametric Engel curves for the reference demographic group (couples with one child) and the second group (couples with two children), together with 80% bootstrap confidence bands at the decile points in the log expenditure distribution for the reference group. In all cases we present Kernel regressions for the Gaussian kernel, using leave-one-out cross-validation methods to automate the choice of bandwidth in each non-parametric regression.[1] Data were trimmed to exclude the top and botton $2\frac{1}{2}$% in each sample. When evaluating bootstrap confidence bands we employ the Smooth Conditional Moment (SCM) method of Gozalo (1997) as a generalization of the Golden Section bootstrap of Härdle and Mammen (1993) to generate 500 bootstrap samples.[2]

[1] Let $\{(\ln x_i, w_{ij})\}_{i=1}^N$ represent a sequence of observations on log expenditure $\ln x_i$ and budget share w_{ij} for the jth good. Further, let $K_h(\cdot) = h^{-1}K(\cdot/h)$ for some symmetric kernel weight function $K(\cdot)$ which integrates to one, given some bandwidth h for which $h \to 0$ and $nh \to \infty$ as $n \to \infty$. We may write the unrestricted Nadaraya–Watson kernel regression estimator of the jth share equation as $\hat{m}_{jh}(\ln x) = (1/N)\Sigma_i W_{ih}(\ln x) \cdot w_{ij}$. where $W_{ih}(\ln x) = K_h(\ln x - \ln x_i)/\hat{f}_h(\ln x)$ and $\hat{f}_h(\ln x) = (1/N)\Sigma_i K_h(\ln x - \ln x_i)$. See Blundell and Duncan (1998) for a survey.

[2] To implement the Gozalo SCM bootstrap method to generate confidence bands for a kernel estimator of the function $w_{ij} = g_j(\ln x_i) + \varepsilon_{ij}$, first form residuals $\hat{\varepsilon}_{ij} = w_{ij} - \hat{m}_{jh}(\ln x)$ from the original sequence of observations $\{(\ln x_i, w_{ij})\}_{i=1}^N$ based on some estimator $\hat{m}_{jh}(\ln x)$ of $g_j(\ln x_i)$. Then evaluate smooth conditional second and third moments at each data point using kernel estimators $\hat{\sigma}^2_{jh}(\ln x) = (1/N)\Sigma_i W_{ih}(\ln x) \cdot \hat{\varepsilon}^2_{ij}$ and $\hat{\mu}^3_{jh}(\ln x) = (1/N)\Sigma_i W_{ih}(\ln x) \cdot \hat{\varepsilon}^3_{ij}$ respectively. Next, draw bootstrap residuals ε^*_{ij} with replacement from a two-point distribution $\hat{F}^{\varepsilon}_{ij}$ defined such that $\Pr(\varepsilon^*_{ij} = a_{ij}) = \gamma_{ij}$ and $\Pr(\varepsilon^*_{ij} = b_{ij}) = 1 - \gamma_{ij}$, where $T_{ij} = [(\hat{\mu}^3_{jh}(\ln x))^2 + 4(\hat{\sigma}^2_{jh}(\ln x))^3]^{1/2}$, $a_{ij} = [\hat{\mu}^3_{jh}(\ln x) - T_{ij}]/(2\hat{\sigma}^2_{jh}(\ln x))$, $b_{ij} = [\hat{\mu}^3_{jh}(\ln x) + T_{ij}]/(2\hat{\sigma}^2_{jh}(\ln x))$ and $\gamma_{ij} = (1/2) \cdot [1 - \hat{\mu}^3_{jh}(\ln x)/T_{ij})]$. Finally, form $w^*_{ij} = \hat{m}_{jh*}(\ln x) + \varepsilon^*_{ij}$ at each stage, where $\hat{m}_{jh*}(\ln x)$ is an oversmoothed kernel estimator of $g_j(\ln x_i)$. Re-estimate $\hat{m}^*_{jh}(\ln x)$ for each bootstrap sample $\{(\ln x_i, w^*_{ij})\}_{i=1}^N$ using the original h, and form empirical quantiles of the bootstrap estimates at a collection of points to generate confidence bands. We use a bandwidth h^* which exceed the cross-validated value by 30%. See Gozalo (1997, pp. 359–363) for a full discussion of the properties of the bootstrap confidence bands, bias and choice of bandwidth.

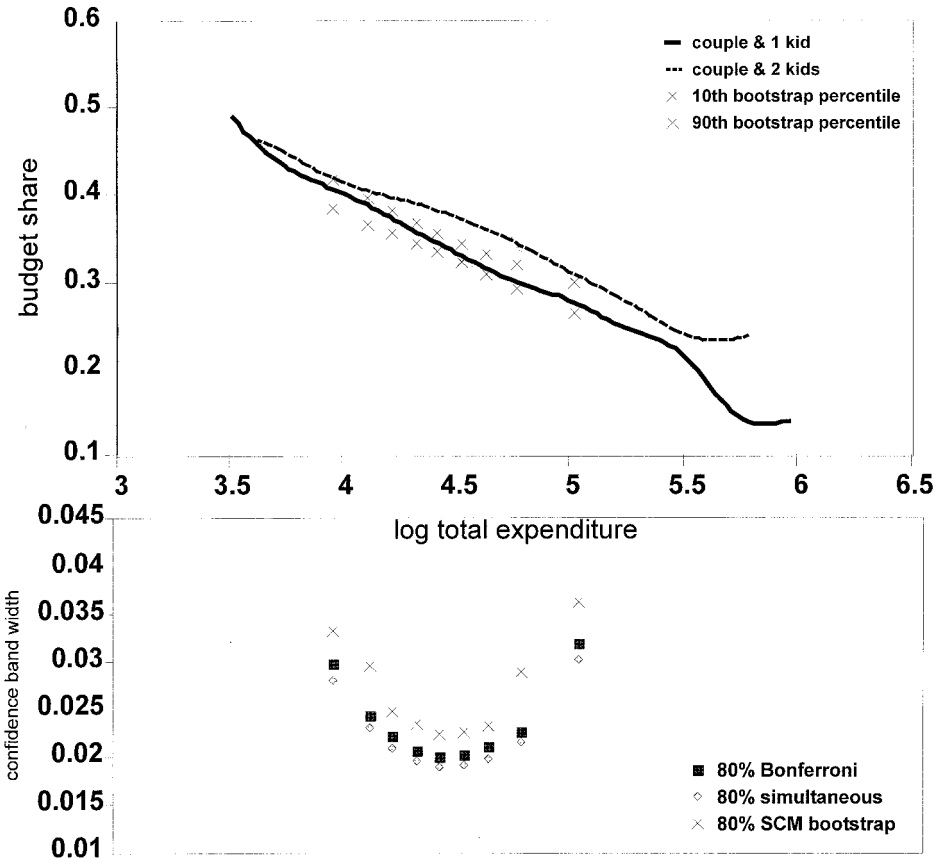

Figure 1. Food Engel curves

In the lower panels of Figures 1 to 6 we present 80% uniform confidence bands evaluated at the nine decile points for the reference demographic group. There are two asymptotic bands: the Bonferroni band that uses asymptotic pointwise bands to construct a joint interval assuming independence and the simultaneous band that accounts for the dependence in the pointwise asymptotic distribution.[3] Finally, we present SCM bootstrap confidence intervals. As expected, the two asymptotic bands yield similar results with the simultaneous band, marked by the diamond, generally slightly narrower reflecting dependence across intervals. The bootstrap bands, marked with a cross, are larger reflecting additional finite sample imprecision.

These regressions would appear to demonstrate that the Working–Leser linear logarithmic (Piglog) formulation is a reasonable approximation for some budget share curves (for example, food and fuel). For other shares, in particular alcohol and other goods, a more non-linear relationship between share and log expenditure is evident. For the alcohol share a quadratic logarithmic share model would seem to fit quite well. These results are consistent with those of

[3] Both formulae can be found in Härdle (1990, section 4.3).

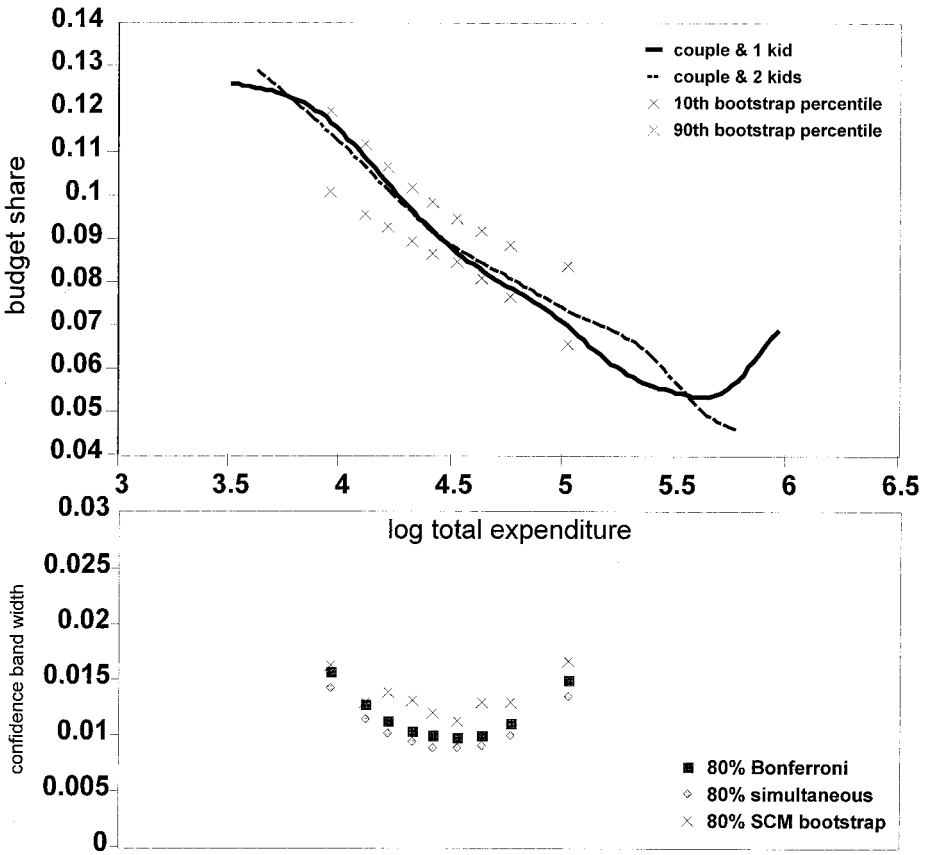

Figure 2. Fuel Engel curves

Banks, Blundell, and Lewbel (1997), although we find less evidence of non-linearity in the clothing Engel curve.[4]

It is interesting to note how similar are the shapes of the Engel curves for our two demographic groups. In Figure 1, for example, we see a broadly parallel shift in the food Engel curve, with couples with two children spending around 4% more of their budget on food thatn couples with a single child (the 'reference' demographic group) at the same (unequivalized) level of total expenditure. For alcohol and transport, on the other hand, Engel curves for couples with two children shift down relative to the reference group (see Figures 4 and 5 respectively).[5] There is no strong evidence of demographic variability in clothing, fuel and other good shares.

[4] The two studies differ in that we analyse the consumption patterns of couples with one or two children, whereas Banks, Blundell, and Lewbel (1997) restrict attention to a more homogeneous group of childless couples. Nevertheless, it is an instructive demonstration of the potential demographic variability in consumption behaviour.

[5] Notice, however, that the two alcohol share curves peak at different log expenditure levels, suggesting that demographic shifts in behaviour combine both horizontal and vertical translations. We shall return to this issue later.

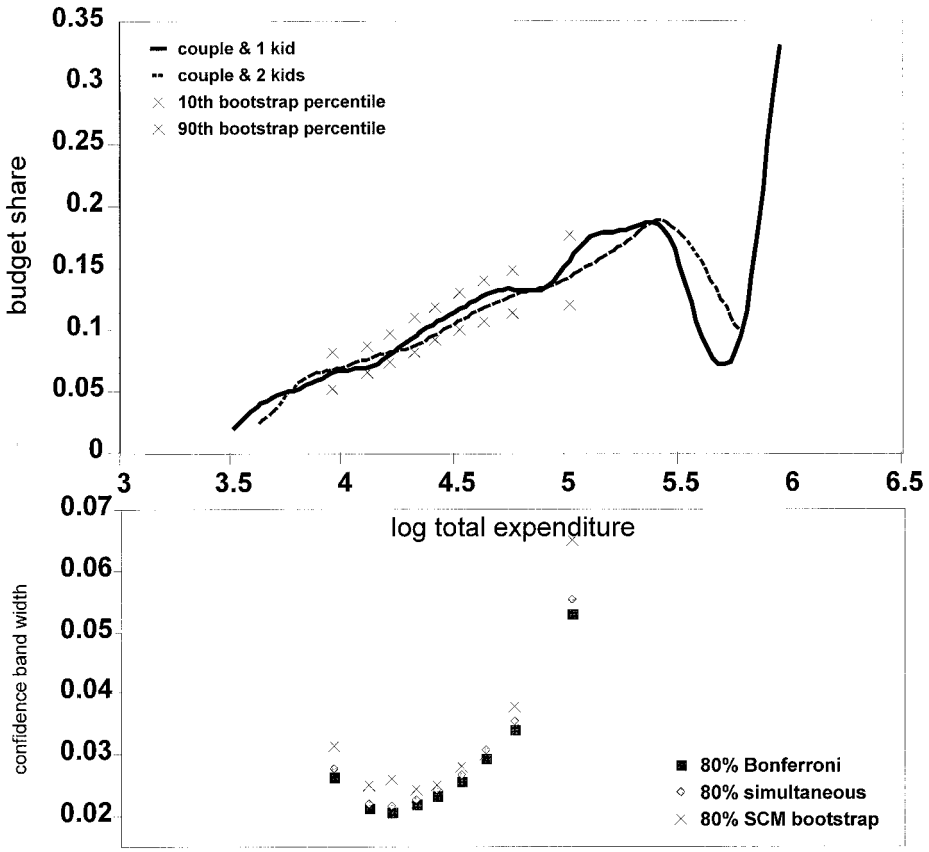

Figure 3. Clothing Engel curves

3. SHAPE RESTRICTIONS AND CONSUMER THEORY

3.1 Semiparametric Specifications for Demographic Composition

The partially linear model

In Engel curve analysis it is important to account for household composition. For example, in the analysis of equivalence scales differences in Engel curves across demographic types are used to construct equivalent income adjustments. In general, knowledge of the way income effects differ across household types is critical in understanding the impact of tax and welfare programmes on expenditure patterns. Any method for incorporating demographic variation must acknowledge this variety in behaviour. One method to account for observed differences in household type is to stratify the sample and implement non-parametric regression within each group. At some point, however, it may be useful to pool across demographic types and to parameterize the way demographic characteristics enter the conditional mean specification. For example, we may be willing to analyse families with and without children separately but may wish to pool our analysis of families with children across different numbers of children in a semiparametric framework.

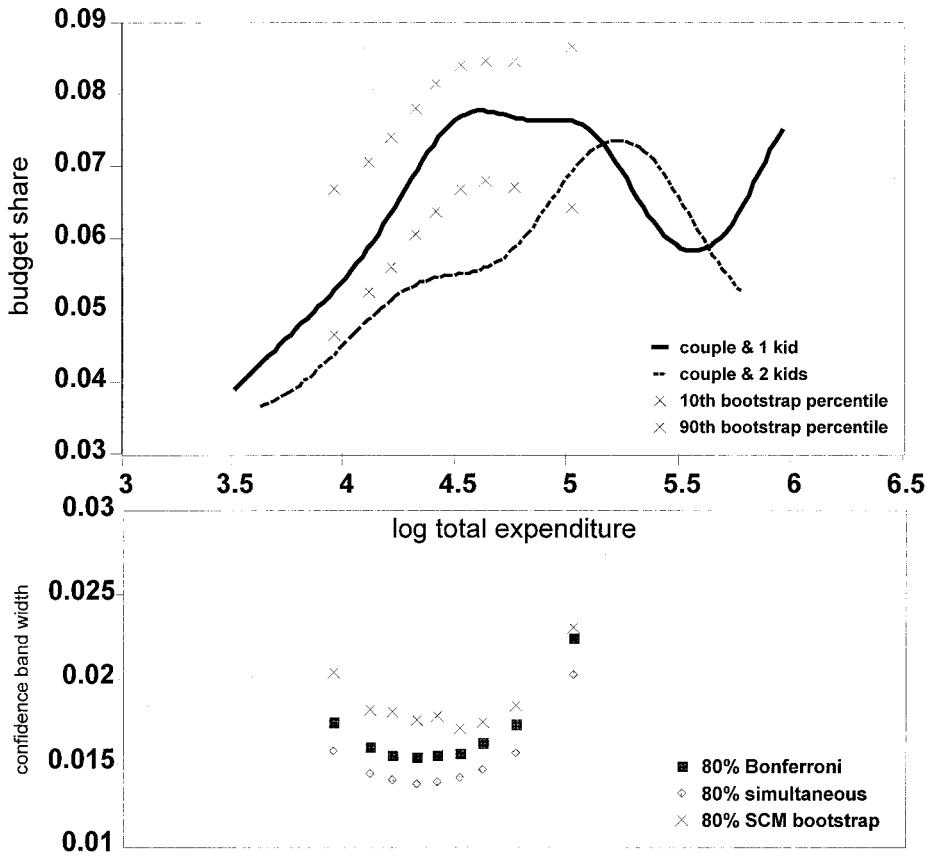

Figure 4. Alcohol Engel curves

A popular approach to semiparametric estimation is to use the following partially linear regression for each expenditure share equation

$$w_{ij} = \alpha_j' z_i + g_j(\ln x_i) + \varepsilon_{ij} \tag{2}$$

in which $\alpha_j' z$ represents a linear index in terms of a finite vector of observable exogenous regressors z_i and unknown parameters α_j. Here we will assume $E(\varepsilon_{ij} \mid z, \ln x) = 0$ and $\mathrm{Var}(\varepsilon_{ij} \mid z, \ln x) = \sigma_j^2(z, \ln x)$. Following Robinson (1988), a simple transformation of the model can be used to give an estimator for α_j. Taking expectations of (2) conditional on $\ln x$, and subtracting from the resulting expression from (2) yields

$$w_{ij} - E(w_{ij} \mid \ln x_i) = \alpha_j'(z_i - E(z_i \mid \ln x_i)) + \varepsilon_{ij} \tag{3}$$

The terms $E(w_{ij} \mid \ln x_i)$ and $E(z_i \mid \ln x_i)$ can be replaced by their non-parametric estimators, denoted $\hat{m}_{jh}^{w}(\ln x)$ and $\hat{m}_{h}^{z}(\ln x)$ respectively, which converge at a slower rate than $\sqrt{n}$. The ordinary least squares estimator for α_j is $\sqrt{n}$ consistent and asymptotically normal.

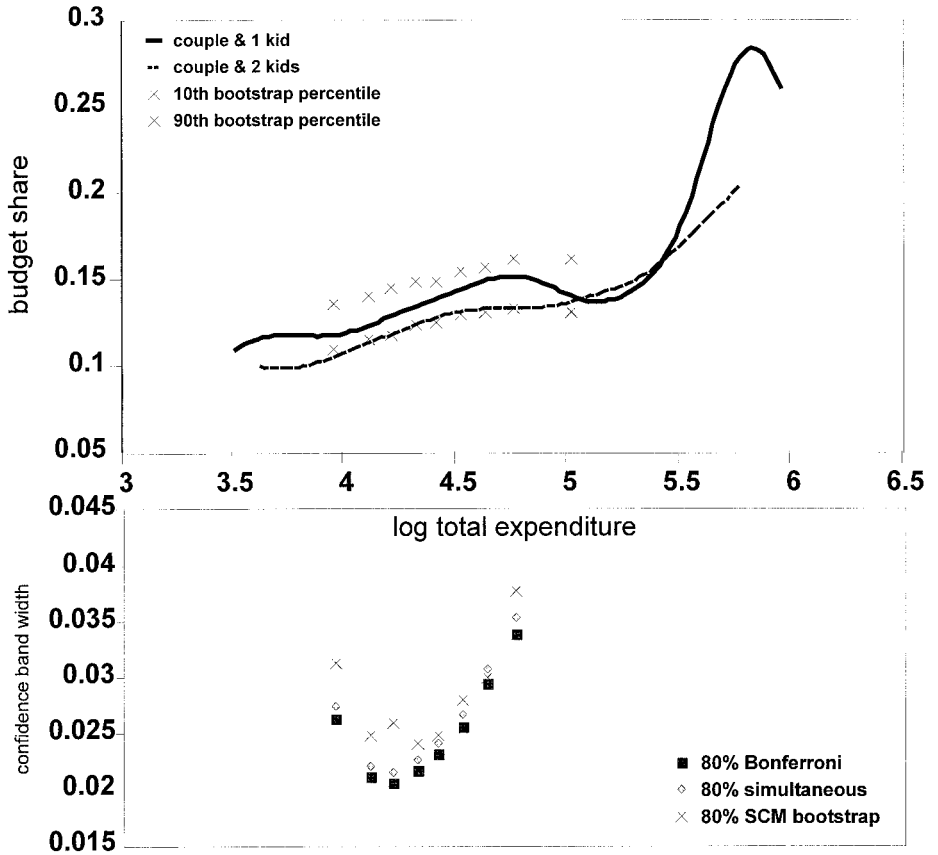

Figure 5. Transport Engel curves

The estimator for $g_j(\ln x)$ is then simply

$$\hat{g}_{jh}(\ln x) = \hat{m}^w_{jh}(\ln x) - \hat{\alpha}'\hat{m}^z_h(\ln x) \tag{4}$$

Since α_j converges at $\sqrt{n}$, but $\hat{m}^w_{jh}(\ln x)$ and $\hat{m}^z_h(\ln x)$ converge at a slower rate, the asymptotic distribution results for $\hat{g}_{jh}(\ln x)$ remain unaffected by estimation of α_j and follows from the distribution of $\hat{m}^w_{jh}(\ln x) - \alpha_j'\hat{m}^z_h(\ln x)$.[6]

Demographic specification and restrictions on consumer preferences

The partially linear model appears to be an attractive method for parsimoniously pooling nonparametric regressions across households with different demographic composition z. However, one may ask under what circumstances equation (2) is consistent with consumer theory. From Shepard's lemma (see Deaton and Muellbauer, 1980b, for example), the budget share equation is

[6] In an interesting recent paper, Heckman *et al.* (1995) show this asymptotic distribution result can provide a poor approximation even in moderately sized samples. They implement bootstrap methods which seem to perform well in Monte Carlo comparisons. These techniques are also used in our application below.

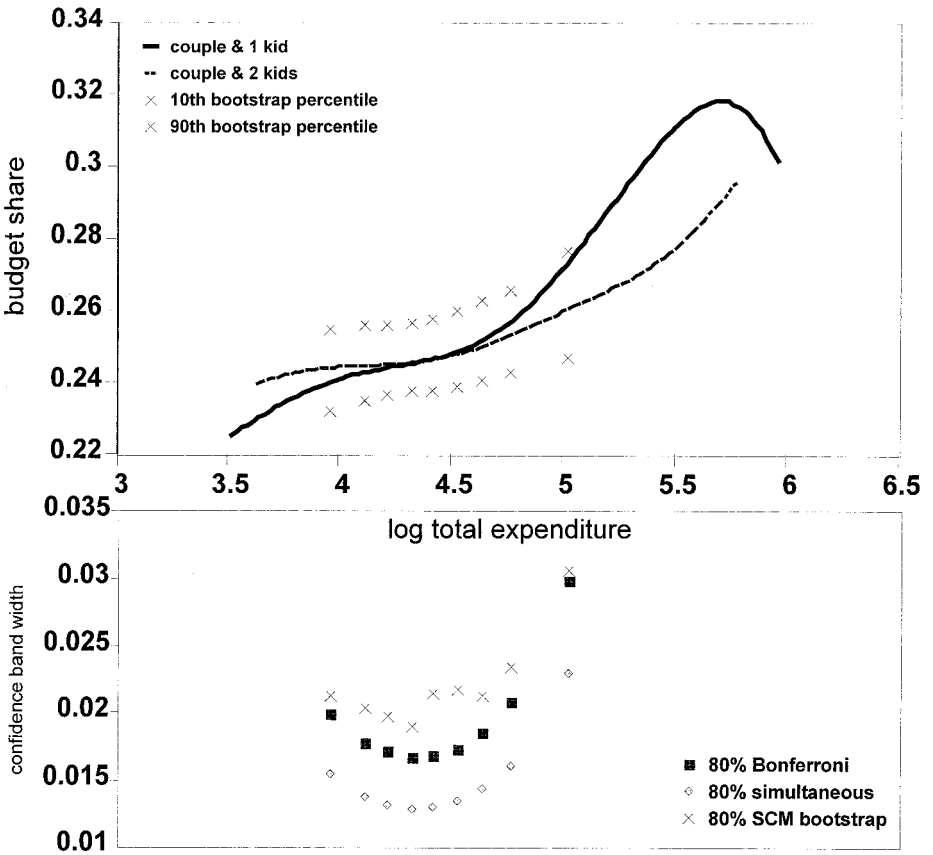

Figure 6. Other goods Engel curves

simply the log price derivative of the consumers expenditure function. Since total expenditure is identical to total costs, this places restrictions on the semiparametric specification. Note that the Engel curve describes the relationship between the expenditures or expenditure shares for a particular value of relative prices.

First we need some definitions: let $P = [P_1, \ldots, P_N]'$ be an N-vector of prices and $p = (\ln(P_1), \ldots, \ln(P_N))'$; define u as utility; and define the cost function, $c(p, u, z)$, as giving the minimum expenditure necessary for a household with characteristics z to achieve utility level u at log-prices p.

Assume that the Engel curve analysis is undertaken in each price regime. That is, the Engel curve takes place at a particular location and time t so that prices can be assumed constant. We write w_{it} as the expenditure share on commodity j for observation t with total budget x_t and the log price N-vector p_t. We suppress the individual household subscript i throughout this discussion of preference restrictions. As above, demographic composition is represented by z_t. In the partially linear model the local average budget share for good j is given by

$$E(w_{jt} \mid z_t, p_t, \ln x_t) = \alpha_j^0(p_t) + \alpha_j^1(p_t)'z_t + g_j(z_t, p_t, \ln x_t) \tag{5}$$

If preferences are consistent with a regular utility-maximizing framework then these shares are consistent with the log price derivatives of the log cost function. From Shepard's Lemma budget shares satisfy

$$w_{jt} = \frac{\partial \ln c(z_t, p_t, u_t)}{\partial \ln p_{jt}}$$

where the cost function $c(z_t, p_t, U_t)$ is some smooth function concave in the n-vector of (unlogged) prices P_t and increasing in u_t. We should also note that $\ln x \equiv \ln c(p, u, z)$.

The following lemma shows that if at least one good has shares that are linear in $\ln x$, for example food shares, then introducing demographics as in the PLM restricts all demands to have shares linear in $\ln x$. Preferences are therefore constrained to be in the Piglog class (see Blundell, Browning and Crawford (1997).

Lemma 3.1 Suppose budget shares have the additive form (which includes the PLM (5)):

$$w_j = \alpha_j^0(p) + \alpha_j^1(p, z) + g_j(\ln x, p) \tag{6}$$

and assume one good has $g_j(\ln x, p) = \ln x - \ln a(p)$ then all goods are PIGLOG.

Proof From the definitions of the cost function and budget shares, equation (6) can be rewritten

$$\frac{\partial \ln c(p, u, z)}{\partial p_j} = \alpha_j^0(p) + \alpha_j^1(p, z) + g_j(\ln c(p, u, z), p)$$

Taking the derivative with respect to the lth log-price p_l for $l \neq j$ we reproduce the Slutsky terms:

$$\frac{\partial^2 \ln c(p, u, z)}{\partial p_j \partial p_l} = \frac{\partial \alpha_j^0(p)}{\partial p_l} + \frac{\partial \alpha_j^1(p, z)}{\partial p_l} + \frac{\partial g_j(\ln c(p, u, z), p)}{\partial p_l} + \frac{\partial g_j(\ln c(p, u, z), p)}{\partial c} w_l \tag{7}$$

Now we consider the partial derivative of equation (7) with respect to z, given p and holding $\ln c \equiv \ln x$ constant. Assuming Slutsky symmetry, we have

$$\frac{\partial^2 \alpha_l^1(p, z)}{\partial p_j \partial z} + \frac{\partial g_l(\ln x, p)}{\partial \ln x} \frac{\partial w_j}{\partial z} = \frac{\partial^2 \alpha_j^1(p, z)}{\partial p_l \partial z} + \frac{\partial g_j(\ln x, p)}{\partial \ln x} \frac{\partial w_l}{\partial z} \text{ for } l \neq j \tag{8}$$

Note that, given the additively separable structure of equation (6) in z and $\ln x$, conditional on $\ln x$ and p, $\partial w_l/\partial z$ and $(\partial^2 \alpha_l^1(p, z))/(\partial p_j \partial z)$ are independent of $\ln x$. The partial derivatives of equation (8) with respect to $\ln x$, conditional on z, of order $s > 1$ are

$$\frac{\partial^s g_l(\ln x, p)}{\partial \ln x^s} \frac{\partial w_j}{\partial z} = \frac{\partial^s g_j(\ln x, p)}{\partial \ln x^s} \frac{\partial w_l}{\partial z} \text{ for } s > 1$$

or for any $\partial w_l/\partial z \neq 0$

$$\frac{\partial^s g_j(\ln x, p)}{\partial \ln x^s} = \frac{\partial w_j}{\partial z} \frac{\dfrac{\partial^s g_l(\ln x, p)}{\partial \ln x^s}}{\dfrac{\partial w_l}{\partial z}} \text{ for all } j \text{ and } s > 1.$$

Note that if $\partial w_j/\partial z = 0$ for any good for which $(\partial^s g_j(\ln x, p))/(\partial \ln x^s) \neq 0$ then $\partial w_l/\partial z = 0$ for all goods with $(\partial^s g_l(\ln x, p))/(\partial \ln x^s) \neq 0$. Finally, if there exists one good l for which

$$\frac{\partial^s g_l(\ln x, p)}{\partial \ln x^s} = 0 \text{ for } s > 1$$

then

$$\frac{\partial^s g_j(\ln x, p)}{\partial \ln x^s} = 0 \text{ for } s > 1$$

and all goods are Piglog.■

The importance of this lemma is in showing the restrictiveness of the partially linear method applied to demographic composition in budget share Engel curves. It says that if for any single good there are restrictions on the shape of the Engel curve, then this will induce restrictions across all goods. We now consider an alternative semiparametric method for pooling across demographic types that relaxes these restrictions.

An extended partially linear model for demographic composition

Consider log-cost functions that are additively separable into a function of prices and demographics and a function of prices and utility as follows:

$$\ln c(p, u, z) = \alpha(p, z) + \ln \bar{c}(p, \psi(u, z)) \tag{9}$$

where $\psi(u, z)$ is a z-specific monotonic transformation of u and z^0 represents the demographic characteristics vector of a reference household type. Normalize $\alpha(p, z)$ so that $\alpha(p, z^0) = 0$ and $\psi(u, z)$ so that $\psi(u, z^0) = \bar{\psi}(u)$. Due to these normalizations, $\ln c(p, u, z^0) = \ln \bar{c}(p, \psi(u, z^0)) = \ln \bar{c}(p, \bar{\psi}(u))$, so that $\ln \bar{c}(p, \bar{\psi}(u))$ is the log-cost function for the reference household type.

To explore this form of preferences further we define $a(p, z) = \exp(\alpha(p, z))$ and $\bar{c}(p, \bar{\psi}(u))$ as the cost function of the reference household type and write the cost function as:

$$c(p, u, z) = a(p, z)\bar{c}(p, \bar{\psi}(u)) \tag{10}$$

Note that $c(p, u, z)$ satisfies homogeneity if and only if $a(p, z)$ is homogeneous of degree zero in prices. Assuming that the cost function of the reference household type has symmetric negative semidefinite Hessian, the cost function for any other household type, $c(p, u, z)$, satisfies the Slutsky conditions only if $(\partial a(p, z)/(\partial P_i \partial P_j)$ is a symmetric negative semidefinite matrix.[7] Thus, if $\bar{c}(p, \psi(u))$ satisfies the Slutsky conditions, then $c(p, u, z)$ satisfies the Slutsky conditions if and only if $a(p, z)$ is weakly concave and homogeneous of degree zero in P.

If $\psi(u, z) = \psi(u)$, then equation (9) reduces to the conditions for the existence of a base-independent equivalence scale discussed by Lewbel (1989) and Blackorby and Donaldson (1993)

[7] The Hessian of this cost function with respect to prices, $(\partial^2 c(p, u, z))/(\partial P_i \partial P_j)$, is given by:

$$\frac{\partial^2 a(p, z)}{\partial P_i \partial P_j} + \frac{\partial a(p, z)}{\partial P_i}\frac{\partial \bar{c}(p, \bar{\psi}(u))}{\partial P_j} + \frac{\partial a(p, z)}{\partial P_j}\frac{\partial \bar{c}(p, \bar{\psi}(u))}{\partial P_i} + \frac{\partial^2 \bar{c}(p, \bar{\psi}(u))}{\partial P_i \partial P_j}$$

Since both $a(p, z)$ and $\bar{c}(p, \psi(u, z))$ are homogeneous, the middle two terms are singular matrices.

and recently explored in the context of semiparametric estimation by Pendakur (1998). In this case, we can rewrite equation (10) as

$$a(p, z) = \frac{c(p, u, z)}{\bar{c}(p, \bar{\psi}(u))} \tag{11}$$

so that $a(p, z)$ is the equivalence scale that relates expenditure needs across household types. Inverting equation (9), we can write the dual indirect utility function as:

$$V(p, \ \ln x, z) = \psi^{-1}(\bar{V}(p, \ln x - \alpha(p, z)), z) \tag{12}$$

where $\bar{V}(p, x)$ is the indirect utility function of the reference household type.

Noting that the monotonic transformation $\psi^{-1}(\cdot)$ does not affect observed share equations, and defining $\bar{w}^i(p, \ln x)$ as the Marshallian share equations of the reference household type, we apply Roy's Identity to get

$$w_j(p, \ln x, z) = \frac{\partial \alpha(p, z)}{\partial p_j} + \bar{w}_j(p, \ln x - \alpha(p, z)) \tag{13}$$

For each commodity, the share equations are related across household types by both a vertical translation, $(\partial \alpha(p, z))/\partial p_i$, which is commodity-specific and a horizontal translation, $\alpha(p, z)$, which is commodity-independent. We will refer to the restrictions given by equation (13) as the Extended Partially Linear Model (EPLM).

This discussion can be summarized in the following lemma which states that although the EPLM restricts the way in which demographics affect demands, unlike the PLM it does not place any further restrictions on preferences.

Lemma 3.2 If budget shares have the EPLM form:

$$w_j = \alpha_j(p, z) + g_j(\ln x - \alpha(z, p)) \tag{14}$$

then, if the reference share equations

$$w_j = g_j(\ln x, p) \tag{15}$$

are consistent with consumer theory and $a(p, z) \equiv \exp(\alpha(z, p))$ is weakly concave and homogeneous of degree zero in P, budget shares given by equation (14) are also consistent with consumer theory.

If we assume that $\{(\partial \alpha(p, z))/\partial p_k\}$ are linear functions of z, then equation (13) has the same vertical translation as the PLM given by equation (6). However, the EPLM also has a horizontal translation in the share equations given by $\alpha(p, z)$. Thus, the EPLM requires that Engel curves exhibit *shape-invariance* across household types.

In the context of non-parametric estimation, where the researcher estimates Engel curves for a single price vector, we do not estimate $\{\partial \alpha(p, z))/\partial p_k\}$ as functionals of $\alpha(p, z)$. Instead, the

researcher assumes that $\{\partial\alpha(p, z))/\partial p_k\}$ are the deratives of $\alpha(p, z)$ at the price vector of estimation. In particular, one could assume that

$$\alpha(p, z) = \phi(z) + \prod_{k=1}^{N} \alpha_k(z) p_k \tag{16}$$

with prices normalized so that $\Pi_{k=1}^{N} \alpha_k(z) p_k = 0$ at the price vector of estimation. Thus the researcher directly estimates $\phi(z)$ and $\{\alpha_k(z)\}$, which correspond to $\alpha(p, z)$ and $\{(\partial\alpha(p, z))/\partial p_k\}$ in equation (13).

If the EPLM holds and the reference share equations are loglinear in total expenditure, then unique $\phi(z)$ and $\{\alpha_k(z)\}$ cannot be recovered. Indeed, Blackorby and Donaldson (1994) show that in this situation, there are an infinite number of $[\alpha(p, z); \{(\partial\alpha(p, z))/\partial p_k\}]$ that would fit the observed share equations. In this case, semiparametric estimation under the PLM would find the unique $\{\alpha_k(z)\}$ under the restriction that $\phi(z) = 0$. Essentially, if reference share equations are loglinear, the PLM can fit the data by mixing the vertical and horizontal translations. With any other shape for the reference household share equations imposing the PML restriction $\phi(z) = 0$ will restrict preferences.[8]

3.2 Shape-invariant Demands and Demographic Composition

Budget shares (equation (14)) are a generalization of the partially linear model. Interestingly, equation (14) has precisely the shape-invariance form found in the extension to the partially linear model considered in the work on pooling non-parametric regression curves by Härdle and Marron (1990), Kneip (1994) and Pinkse and Robinson (1995).

Suppose z is binary, $z = \{0, 1\}$, and consider data drawn from a single price regime. This is consistent with our empirical analysis where we consider the differences between demands in a single period for couples with one child and couples with two. In this binary case, we normalize on the reference type $z = 0$, so $\alpha_j^0 = 0$ for all j and $\phi^0 = 0$. Then, denote $\alpha_j = \alpha_j^1$ and $\phi = \phi^1$, so that α_j are scaler parameters for each share equation and ϕ is a single parameter common to all equations. To estimate we use an approach to pooling in non-parametric regression due to Pinkse and Robinson (1995) which adapts the idea of Härdle and Marron (1990). Suppose also that the unrestricted non-parametric regression has been estimated separately on N^z datapoints for each subgroup, $z = \{0, 1\}$. For each good j define

$$\hat{f}^z(\ln x) = \frac{1}{N^z} \sum_{i|_{z=z}} K_h(\ln x - \ln x_i) \tag{17}$$

[8] Gozalo (1997) and Pendakur (1998) have estimated semiparametric demand systems similar to the EPLM, both in the context of investigating household equivalence scales. Both papers test the shape-invariance restrictions given by the EPLM on household Engel curves. Gozalo finds that if the food price elasticity of the equivalence scale is restricted to be zero, $\alpha_{\text{food}}(z) = 0\ \forall z$, then shape-invariance is rejected in the data. This amounts to testing the EPLM with all vertical translations forced to zero. Pendakur (1998) tests the EPLM allowing for both vertical and horizontal translations and finds some support in the data for the EPLM.

and

$$\hat{r}_j^z(\ln x) = \frac{1}{N^z} \sum_{i|_{z=z}} K_h(\ln x - \ln x_i) w_{ij} \tag{18}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ for some symmetric kernel weight function $K(\cdot)$ which integrates to one, and for some bandwidth h for which $h \to 0$ and $nh \to \infty$ as $n \to \infty$. We can write the Nadaraya–Watson kernel regression estimates for each subgroup as $\hat{m}_j^z(\ln x) = \hat{r}_j^z(\ln x)/\hat{f}^z(\ln x)$ for $z = 0, 1$. Here, $\hat{r}_j^z(\ln x)$ are the convolved share data for the two types and $\hat{f}^z(\ln x)$ are kernel density estimates for the two types. The dependence of these functions on the bandwidth, h, is suppressed.

The restrictions for the EPLM may be written

$$\hat{m}_j^1(\ln x) = \alpha_j + \hat{m}_j^0(\ln x - \phi) \tag{19}$$

or

$$\hat{m}_j^1(\ln x) - \hat{m}_j^0(\ln x - \phi) - \alpha_j = 0 \tag{20}$$

Pinkse and Robinson (1995) suggest multiplying equation (20) by $\hat{f}^1(\ln x)\hat{f}^0(\ln x - \phi)$ to obtain

$$\hat{f}^0(\ln x - \phi)\hat{r}_j^1(\ln x) - \hat{f}^1(\ln x)\hat{r}_j^0(\ln x - \phi) - \hat{f}^1(\ln x)\hat{f}^0(\ln x - \phi)\alpha_j = 0 \tag{21}$$

Since (21) does not contain the division operator present in $\hat{m}_j^z(\ln x)$, Pinkse and Robinson (1995) are able to establish asymptotic convergence results for an estimator of ϕ and $\{\alpha_j\}$ that results from minimizing the integrated squared loss function

$$L(\phi, \{\alpha_j\}) = \sum_{j=1}^{n} \int_{\ln \underline{x}}^{\ln \bar{x}} (\Lambda_j(\ln x; \phi, \alpha_j))^2 \Omega_j \cdot \mathrm{d} \ln x \tag{22}$$

where $\ln \underline{x}$ and $\ln \bar{x}$ are integration limits on the log of expenditure,

$$\Lambda_j(\ln x; \phi, \alpha_j) = \hat{f}^1(\ln x)\hat{f}^0(\ln x - \phi)(\hat{m}_j^1(\ln x) - \hat{m}_j^0(\ln x - \phi) - \alpha_j) \tag{23}$$

and Ω_j is an equation-specific weighting function.[9]

[9] There are a number of practical difficulties in the minimisation of equation (22). First, the loss function approaches zero for large negative or positive ϕ, since in either case the product $\hat{f}^1(\ln x)\hat{f}^0(\ln x - \phi)$ becomes arbitrarily small. We therefore implement a restricted gridsearch over a reasonable range for ϕ in order to establish a value at which the loss function attains the relevant local minimum.
The second problem relates to the evaluation of the kernel density and regression terms $\hat{f}^0(\ln x - \phi)$ and $\hat{m}_j^0(\ln x - \phi)$ in equation (22) as one approaches the boundary of the (common) support for $\ln x$. The practical implementation of the Pinkse and Robinson (1995) estimator requires an appropriate choice for the integration limits $\ln \underline{x}$ and $\ln \bar{x}$ on log total expenditure such that $\hat{f}^0(\ln x - \phi)$ exists over the (restricted) range $\ln x \in \{\ln \underline{x}, \ln \bar{x}\}$ given observed data.

4. SEMIPARAMETRIC ESTIMATION RESULTS

4.1 Specification Testing and Endogeneity

How well does the quadratic logarithmic specification, underlying the QUAIDS, fit in comparison with these semiparametric specifications? In this section we use the semiparametric regression models as an alternative against which to test a parametric quadratic logarithmic null. Convenient goodness of fit tests have been proposed by Härdle and Mammen (1993) and extended in Aït-Sahalia, Bickel, and Stoker (1994). These studies derive asymptotically normal statistics for the comparison between a non-parametric estimate $\hat{g}_{jh}(\ln x_i)$ and some parametric estimate $\delta(\ln x_i, \hat{\beta}_j)$ of a regression curve based on a simple squared error goodness of fit statistic

$$\hat{\Gamma}_j = \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{jh}(\ln x_i) - \delta(\ln x_i, \hat{\beta}_j))^2 w(\hat{f}_h(\ln x_i)) \tag{24}$$

a linear transformation of which is shown to converge at rate $nh^{1/2}$ to a limiting normal distribution with estimable asymptotic bias.[10]

Our proposed test for shape invariance in the semiparametric model adapts Pendakur (1998) by comparing the minimized value of the loss function (22) with that which we would expect under the null of shape invariance. We extend our use of the SCM bootstrap algorithm along the lines suggested by Gozalo (1997) to generate the empirical distribution of the loss function under the extended partially linear null.[11] Bootstrap p-values are presented alongside the value of the minimized loss function for each share equation.

To adjust for endogeneity we adapt the popular augmented regression technique (see Holly and Sargan, 1982, for example) to the semiparametric framework. In particular, suppose $\ln x$ is endogenous in model (1) in the sense that

$$E(\varepsilon_j \mid \ln x) \neq 0 \text{ or } E(w_j \mid \ln x_i) \neq g_j(\ln x) \tag{25}$$

In this case the non-parametric estimator will not be consistent for the function of interest. It will not provide the appropriate counterfactual: how do expenditure share patterns change for some given change in total expenditure? However, suppose there exists a variable y such that

$$\ln x = y \cdot \pi + v \text{ with } E(v \mid y) = 0 \tag{26}$$

Moreover, assume the following linear conditional model holds:

$$w_j = g_j(\ln x) + v \cdot \rho_j + \varepsilon_j \tag{27}$$

[10] An alternative approach by Zheng (1996) uses the kernel method to construct a moment condition which can be used to distinguish the parametric null from the non-parametric alternative. A test proposed by Ellison and Ellison (1992) has a structure almost identical to that of Zheng (1996), and differs only in the form of the variance estimator.

[11] Specifically, the sequence of SCM bootstrap samples used to simulate the empirical distribution of equation (22) under the null derives from resampled budget shares $w_{ij}^* = \hat{\alpha}_j + \hat{m}_{jh}(\ln x_i - \hat{\phi} z_i) + \varepsilon_{ij}^*$ where ε_{ij}^* are defined as for the SCM bootstrap confidence band algorithm. Re-estimating the semiparametric $\hat{\alpha}_j^* + \hat{m}_{jh}^*(\ln x_i - \hat{\phi}^* z_i)$ for each bootstrap sample $\{(\ln x_i, w_{ij}^*)\}_{i=1}^N$ using the original h enables us to build an empirical distribution for the loss function under the null. In practice this is computationally intensive procedure, given that the loss function needs to be re-minimized for each bootstrap sample. We base empirical p-values for the loss function on 500 bootstrap samples.

with

$$E(\varepsilon_j \mid \ln x) = 0 \tag{28}$$

Note that

$$w_j - E(w_j \mid \ln x) = (v - E(v \mid \ln x))\rho_j + \varepsilon_j \tag{29}$$

The estimator of $g_j(\ln x)$ is given by

$$\hat{g}_{jh}(\ln x) = \hat{m}^w_{jh}(\ln x) - \hat{m}^v_h(\ln x)\hat{\rho}_j \tag{30}$$

In place of the unobservable error component v we use the first-stage residuals

$$\hat{v} = x - y\hat{\pi} \tag{31}$$

where $\hat{\pi}$ is the least squares estimator of π. Since $\hat{\pi}$ and $\hat{\rho}_j$ converge at $\sqrt{n}$ the asymptotic distribution for $\hat{g}_{jh}(x)$ follows the distribution of $\hat{m}^w_{jh}(\ln x) - \hat{m}^v_h(\ln x)\rho_j$. Moreover, a test of the exogeneity null, $H_0 : \rho_j = 0$, can be constructed from this least squares regression.

Newey, Powell, and Vella (1995) have developed a generalization of this idea for triangular simultaneous equation systems of the type considered here. They adopt a series approach to the estimation of the regression of w_j on $\ln x$ and v. This generalizes the form of equation (27) and allows an assessment of the additive structure. They also use a non-parametric regression for the reduced form in place of the linear model (26).

In our application we consider extending model (27) along the lines suggested by Newey, Powell, and Vella (1995). This is done by including higher-order terms in the residuals v and then testing the partially linear specification (27) against this more general additive recursive alternative. The first-stage residual $\hat{v}$ in (27) is calculated using the log of disposable income and is used as the excluded instrumental variable.

4.2 Empirical Results

We report a range of semiparametric estimates for the parameters of share equations which, in their most general form, may be written as

$$w_j = \alpha_j \cdot z + g_j(\ln x - \phi \cdot z) + v \cdot \rho_j + \varepsilon_j \tag{32}$$

The first column in each of the following tables contains results for a simple regression of budget share on log expenditure with no semiparametric controls (that is, $\phi = 0$ and $\alpha_j = \rho_j = 0$ for all j). Relative to this benchmark, the second column reports results for a model which adjusts for the number of children in the household ($\alpha_j \neq 0$ for all j) using the partially linear framework of Robinson (1988). The model that controls for demographics and endogeneity ($\alpha_j \neq 0$, $\rho_j \neq 0$) makes up our third specification. The final two specifications relate to the shape-invariant generalizations to the basic Robinson-type model. The fourth model allows for scale shifts in log expenditure by demographic type ($\phi \neq 0$, $\alpha_j \neq 0$ for all j) using the estimation method of Pinkse and Robinson (1995) and the fifth in addition introduces controls for endogeneity ($\phi \neq 0, \alpha_j \neq 0, \rho_j \neq 0$ for all j). Estimation results for these five different semiparametric specifications are presented in Tables II to VII for each of the six share aggregates.

Table II. Non-parametric and semiparametric estimates: food Engel curves

	$\phi = 0$			$\phi = 0{\cdot}2590$ (0·0809)	
	1 No corrections	2 Demographics	3 Demographics and endogeneity	4 Demographics	5 Demographics and endogeneity
$\hat{\beta}_j^{\text{is}}$	−0·1280	−0·1346	−0·1175	−0·1234	−0·1046
	(0·0122)	(0·0117)	(0·0117)	(0·0104)	(0·0105)
$\hat{\beta}_j^{\text{ols}}$	−0·1288	−0·1348	−0·1178	−0·1267	−0·1081
	(0·0083)	(0·0081)	(0·0081)	(0·0084)	(0·0083)
$\hat{\alpha}_j$		0·0338	0·0323	0·0281	0·0273
		(0·0051)	(0·0052)	(0·0048)	(0·0048)
$\hat{\rho}_j$			−0·0242		−0·0276
			(0·0134)		(0·0131)
Loss				0·2295	
				[0·476]	
$\chi_v^2(1)$			2·680		0·947
			[0·102]		[0·330]
H_0: linear parametric form					
$\chi^2_{\text{abs}}(1)$	0·422	0·757	0·881	0·798	0·694
	[0·516]	[0·384]	[0·348]	[0·372]	[0·405]
H_0: quadratic parametric form					
$\chi^2_{\text{abs}}(1)$	1·192	0·853	1·103	0·005	0·004
	[0·275]	[0·356]	[0·294]	[0·944]	[0·950]

Notes: Here and in Tables III to VII data are drawn from the 1980–82 Family Expenditure Surveys. Standard errors in ( ) parentheses and *p*-values in [ ] parentheses. Non-parametric estimates based on a Gaussian kernel with bandwidths chosen by cross-validation (cf. Härdle, 1990). Average derivatives $\hat{\beta}_{\text{is}}$ are indirect slope estimates (cf. Stoker, 1991) for the non-parametric function $g_j(\cdot)$ in equation (32). For cross-validation and ADE calculations, data are trimmed to exclude the smallest 2% of estimated densities. All estimates and specification tests are generated using the GAUSS-based software package NP-REG (see Duncan and Jones, 1992).

In these tables $\hat{\beta}_j^{\text{is}}$ refers to the average derivative (indirect slope) estimates for non-parametric function $g_j(\cdot)$ in equation (32).[12] By way of comparison $\hat{\beta}_j^{\text{ols}}$ refers to the simple ordinary least squares estimate of the slope coefficient under the simple parametric assumption that $g_j(\ln x - \phi \cdot z) = \hat{\beta}_j \cdot (\ln x - \phi \cdot z)$. The shape-invariant transformation (14) has two parameters for each share equation; the scaling parameter ϕ in the term $\ln x - \phi z$ and an intercept parameter α_j. For the latter two models in each of Tables II to VII we estimate the parameters $(\phi, \{\alpha_j\})$ through minimization of equation (22).[13]

Shape-invariant parameter estimates

We estimate the scale parameter ϕ common to all six share equations to be 0·2590 with an SCM bootstrap standard error of 0·0809, giving an estimated equivalence scale of 1·295 for

[12] See Stoker (1991) for a full discussion of various Average Derivative estimators and their properties.

[13] In practice we use sequential gridsearch methods to estimate the scale parameter ϕ and the shift parameters $\{\alpha_j\}$. Initial values for $\{\alpha_j\}$ are estimated using Robinson's (1988) method. Conditional on $\{\hat{\alpha}_j\}$ we then gridsearch the loss function to estimate ϕ. This process is then repeated until convergence is achieved. We generate bootstrap standard errors for $\hat{\phi}$ through repetition of this gridsearch process for 500 bootstrap samples, each generated using the SCM algorithm of Gozalo (1997).

Table III. Non-parametric and semiparametric estimates: fuel Engel curves

	$\phi = 0$			$\phi = 0{\cdot}2590$ (0·0809)	
	1 No corrections	2 Demographics	3 Demographics and endogeneity	4 Demographics	5 Demographics and endogeneity
β_j^{is}	−0·0516	−0·0513	−0·0235	−0·0472	−0·0206
	(0·0045)	(0·0045)	(0·0046)	(0·0042)	(0·0043)
$\hat{\beta}_j^{ols}$	−0·0493	−0·0491	−0·0214	−0·0463	−0·0199
	(0·0044)	(0·0044)	(0·0044)	(0·0043)	(0·0043)
$\hat{\alpha}_j$		0·0017	−0·0004	−0·0013	−0·0022
		(0·0026)	(0·0027)	(0·0025)	(0·0026)
$\hat{\rho}_j$			−0·0350		−0·0382
			(0·0068)		(0·0068)
Loss				0·0512	
				[0·298]	
$\chi_v^2(1)$			2·582		3·571
			[0·108]		[0·059]
H_0: linear parametric form					
$\chi^2_{abs}(1)$	0·416	0·379	0·686	0·197	1·319
	[0·519]	[0·538]	[0·407]	[0·657]	[0·251]
H_0: quadratic parametric form					
$\chi^2_{abs}(1)$	0·092	0·095	0·022	0·028	0·656
	[0·761]	[0·758]	[0·882]	[0·868]	[0·418]

Notes: Here and in Tables III to VII data are drawn from the 1980–82 Family Expenditure Surveys. Standard errors in ( ) parentheses and *p*-values in [ ] parentheses. Non-parametric estimates based on a Gaussian kernel with bandwidths chosen by cross-validation (cf. Härdle, 1990). Average derivatives $\hat{\beta}_{is}$ are indirect slope estimates (cf. Stoker, 1991) for the non-parametric function $g_j(\cdot)$ in equation (32). For cross-validation and ADE calculations, data are trimmed to exclude the smallest 2% of estimated densities. All estimates and specification tests are generated using the GAUSS-based software package NP-REG (see Duncan and Jones, 1992).

couples with two children compared with our reference group. This accords quite closely with estimates reported in Pendakur (1998) for a range of US and UK studies. The parameters $\hat{\alpha}_j$ specific to each share equation are reported in the tables. Having accounted for the scale parameter ϕ, we find significant shift parameters for food (positive), alcohol and transport (both negative), confirming the initial graphical evidence in Figures 1 to 6.

Average derivative estimates

Compared with the first specification, the average slope of the food Engel curve, estimated by the indirect average derivative $\hat{\beta}_j^{is}$ in Table II, becomes more negative when controlled for household size using the shape-invariant model but less so once the correction for endogeneity is included. Notice how the average marginal effect of log expenditure on food share reduces when one controls more fully for demographic variability using the Extended Partially Linear Model. Notice also how the inclusion of the scale parameter impacts on the magnitude of the estimate of the shift parameter α_j. In particular we see a lower value for $\hat{\alpha}_j$ in the food share equation once log expenditure has been equivalized for household size.

Table IV. Non-parametric and semiparametric estimates: clothing Engel curves

	$\phi = 0$			$\phi = 0{\cdot}2590$ (0·0809)	
	1 No corrections	2 Demographics	3 Demographics and endogeneity	4 Demographics	5 Demographics and endogeneity
$\hat{\beta}_j^{\text{is}}$	0·0910	0·0914	0·0518	0·0855	0·0473
	(0·0083)	(0·0083)	(0·0081)	(0·0087)	(0·0083)
$\hat{\beta}_j^{\text{ols}}$	0·0882	0·0885	0·0493	0·0864	0·0485
	(0·0083)	(0·0083)	(0·0082)	(0·0083)	(0·0082)
$\hat{\alpha}_j$		−0·0049	−0·0014	−0·0018	0·0004
		(0·0047)	(0·0049)	(0·0045)	(0·0046)
$\hat{\rho}_j$			0·0527		0·0555
			(0·0129)		(0·0127)
Loss				0·0965	
				[0·096]	
$\chi_v^2(1)$			5·919		3·591
			[0·015]		[0·058]
H_0: linear parametric form					
$\chi^2_{\text{abs}}(1)$	0·798	0·812	0·774	0·711	0·973
	[0·372]	[0·368]	[0·379]	[0·399]	[0·324]
H_0: quadratic parametric form					
$\chi^2_{\text{abs}}(1)$	0·600	0·611	0·649	0·844	1·260
	[0·439]	[0·435]	[0·421]	[0·358]	[0·262]

Notes: Here and in Tables III to VII data are drawn from the 1980–82 Family Expenditure Surveys. Standard errors in ( ) parentheses and *p*-values in [ ] parentheses. Non-parametric estimates based on a Gaussian kernel with bandwidths chosen by cross-validation (cf. Härdle, 1990). Average derivatives $\hat{\beta}_{\text{is}}$ are indirect slope estimates (cf. Stoker, 1991) for the non-parametric function $g_j(\cdot)$ in equation (32). For cross-validation and ADE calculations, data are trimmed to exclude the smallest 2% of estimated densities. All estimates and specification tests are generated using the GAUSS-based software package NP-REG (see Duncan and Jones, 1992).

Specification test results

We report empirical *p*-values (denoted *p*[*Loss*]) for tests of the null of shape invariance against a fully non-parametric alternative for all share equations. We are unable to reject shape invariance in all cases, even for share relationships (e.g. alcohol) which are less obviously shape invariant from casual graphical examination. We also report tests of the linear (Working–Leser) and quadratic logarithmic specifications against the semiparametric alternatives for each semiparametric model in Tables II–VII. For the food share in Table II, in all specifications, we are unable to reject linearity. In contrast, for alcohol share, the Piglog of Working–Leser form is strongly rejected. In line with Blundell, Pashardes, and Weber (1993), the quadratic logarithmic specification is not rejected by the data. This result is maintained even after controlling for demographic variation and the endogeneity of total expenditure.

We find the correction for endogeneity of log total expenditure to be important in most share equations, most notably food, fuel, clothing and alcohol share. The χ_v^2 statistic refers to a one degree of freedom test of the conditionally linear endogeneity correction (27) against the inclusion of higher terms in *v*. Here we simply consider an alternative that includes a second-order residual. There is little evidence against the conditionally linear correction.

Table V. Non-parametric and semiparametric estimates: alcohol Engel curves

	$\phi = 0$			$\phi = 0{\cdot}2590$ (0·0809)	
	1 No corrections	2 Demographics	3 Demographics and endogeneity	4 Demographics	5 Demographics and endogeneity
$\hat{\beta}_j^{\text{is}}$	0·0243	0·0264	0·0119	0·0236	0·0076
	(0·0037)	(0·0037)	(0·0037)	(0·0035)	(0·0036)
$\hat{\beta}_j^{\text{ols}}$	0·0215	0·0231	0·0087	0·0191	0·0030
	(0·0050)	(0·0050)	(0·0049)	(0·0047)	(0·0047)
$\hat{\alpha}_j$		−0·0139	−0·0127	−0·0121	−0·0115
		(0·0034)	(0·0034)	(0·0032)	(0·0033)
$\hat{\rho}_j$			0·0198		0·0230
			(0·0088)		(0·0086)
Loss				0·1483	
				[0·144]	
$\chi_v^2(1)$			0·726		0·941
			[0·394]		[0·332]
H_0: linear parametric form					
$\chi^2_{\text{abs}}(1)$	5·146	8·621	8·167	7·887	9·567
	[0·023]	[0·003]	[0·004]	[0·005]	[0·002]
H_0: quadratic parametric form					
$\chi^2_{\text{abs}}(1)$	0·044	0·127	0·083	0·159	0·397
	[0·833]	[0·721]	[0·773]	[0·690]	[0·529]

Notes: Here and in Tables III to VII data are drawn from the 1980–82 Family Expenditure Surveys. Standard errors in ( ) parentheses and *p*-values in [ ] parentheses. Non-parametric estimates based on a Gaussian kernel with bandwidths chosen by cross-validation (cf. Härdle, 1990). Average derivatives $\hat{\beta}_{\text{is}}$ are indirect slope estimates (cf. Stoker, 1991) for the non-parametric function $g_j(\cdot)$ in equation (32). For cross-validation and ADE calculations, data are trimmed to exclude the smallest 2% of estimated densities. All estimates and specification tests are generated using the GAUSS-based software package NP-REG (see Duncan and Jones, 1992).

4.3 A Graphical Analysis of Shape Invariance

For a graphical comparison of the alternative specifications we consider the shape-invariant restricted models without endogeneity correction. These correspond to the fourth columns in Tables II–VII and are presented graphically in Figures 7 to 12. The solid line is the reference curve (for couples with one child). The hashed line is the unrestricted equivalent kernel regression curve for families with two children. The dotted lines are the shape-invariant curves using estimates from Tables II–VII. Note that shape-invariant and unrestricted curves are, in most cases, quite comparable, and consistent with the bootstrap specification tests of shape invariance reported earlier.

5. CONCLUSIONS

This paper has been concerned with investigating the 'shape' of consumer preferences using semiparametric methods. By choosing consumers from a point in time and location we have focused on the Engel curve relationship. As a baseline specification we have worked with the Working–Leser or Piglog specification in which budget shares are expressed in terms of log total

Table VI. Non-parametric and semiparametric estimates: transport Engel curves

	$\phi = 0$			$\phi = 0{\cdot}2590$ (0·0809)	
	1 No corrections	2 Demographics	3 Demographics and endogeneity	4 Demographics	5 Demographics and endogeneity
$\hat{\beta}_j^{\text{is}}$	0·0328	0·0349	0·0188	0·0338	0·0180
	(0·0071)	(0·0072)	(0·0072)	(0·0068)	(0·0068)
$\hat{\beta}_j^{\text{ols}}$	0·0295	0·0314	0·0152	0·0302	0·0144
	(0·0088)	(0·0088)	(0·0088)	(0·0087)	(0·0087)
$\hat{\alpha}_j$		−0·0124	−0·0111	−0·0100	−0·0094
		(0·0055)	(0·0056)	(0·0053)	(0·0053)
$\hat{\rho}_j$			0·0204		0·0228
			(0·0144)		(0·0142)
Loss				0·1360	
				[0·242]	
$\chi_v^2(1)$			0·0122		0·231
			[0·912]		[0·631]
H_0: linear parametric form					
$\chi^2_{\text{abs}}(1)$	0·431	0·672	0·595	0·057	0·083
	[0·511]	[0·413]	[0·441]	[0·811]	[0·773]
H_0: quadratic parametric form					
$\chi^2_{\text{abs}}(1)$	0·719	0·707	0·801	0·826	0·935
	[0·396]	[0·400]	[0·371]	[0·364]	[0·334]

Notes: Here and in Tables III to VII data are drawn from the 1980–82 Family Expenditure Surveys. Standard errors in ( ) parentheses and *p*-values in [ ] parentheses. Non-parametric estimates based on a Gaussian kernel with bandwidths chosen by cross-validation (cf. Härdle, 1990). Average derivatives $\hat{\beta}_{\text{is}}$ are indirect slope estimates (cf. Stoker, 1991) for the non-parametric function $g_j(\cdot)$ in equation (32). For cross-validation and ADE calculations, data are trimmed to exclude the smallest 2% of estimated densities. All estimates and specification tests are generated using the GAUSS-based software package NP-REG (see Duncan and Jones, 1992).

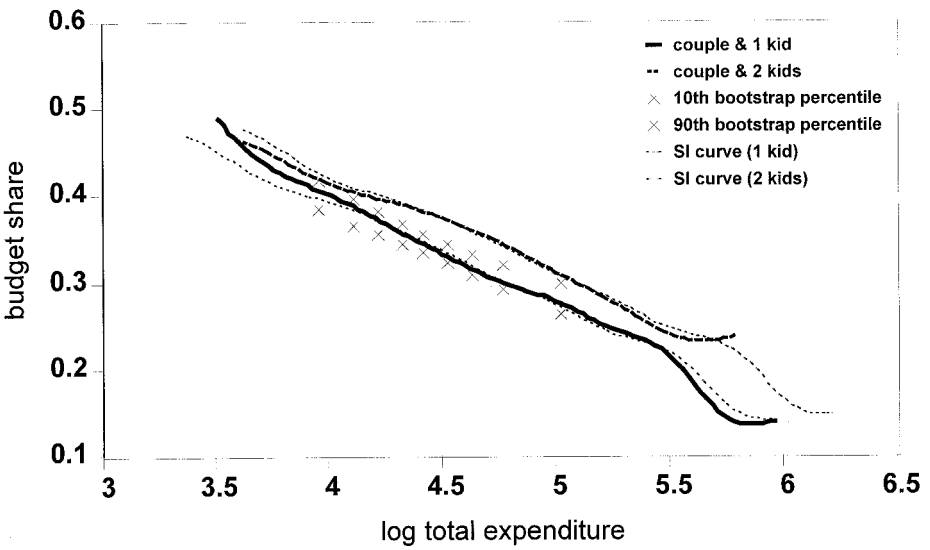

Figure 7. Shape-invariant transformation: food share

Table VII. Non-parametric and semiparametric estimates: other goods Engel curves

	$\phi = 0$			$\phi = 0{\cdot}2590$ (0·0809)	
	1 No corrections	2 Demographics	3 Demographics and endogeneity	4 Demographics	5 Demographics and endogeneity
$\hat{\beta}_j^{\text{is}}$	0·0329	0·0333	0·0468	0·0313	0·0447
	(0·0097)	(0·0097)	(0·0097)	(0·0087)	(0·0088)
$\hat{\beta}_j^{\text{ols}}$	0·0358	0·0362	0·0497	0·0337	0·0471
	(0·0091)	(0·0091)	(0·0091)	(0·0090)	(0·0090)
$\hat{\alpha}_j$		−0·0035	−0·0043	−0·0016	−0·0018
		(0·0054)	(0·0055)	(0·0052)	(0·0053)
$\hat{\rho}_j$			−0·0149		−0·0193
			(0·0126)		(0·0125)
Loss				0·0043	
				[0·164]	
$\chi_v^2(1)$			0·0569		0·506
			[0·811]		[0·477]
H_0: linear parametric form					
$\chi^2_{\text{abs}}(1)$	1·933	1·901	4·336	0·591	1·404
	[0·164]	[0·168]	[0·037]	[0·442]	[0·236]
H_0: quadratic parametric form					
$\chi^2_{\text{abs}}(1)$	0·005	0·004	0·140	0·364	0·067
	[0·943]	[0·947]	[0·708]	[0·546]	[0·796]

Notes: Here and in Tables III to VII data are drawn from the 1980–82 Family Expenditure Surveys. Standard errors in ( ) parentheses and *p*-values in [ ] parentheses. Non-parametric estimates based on a Gaussian kernel with bandwidths chosen by cross-validation (cf. Härdle, 1990). Average derivatives $\hat{\beta}_{\text{is}}$ are indirect slope estimates (cf. Stoker, 1991) for the non-parametric function $g_j(\cdot)$ in equation (32). For cross-validation and ADE calculations, data are trimmed to exclude the smallest 2% of estimated densities. All estimates and specification tests are generated using the GAUSS-based software package NP-REG (see Duncan and Jones, 1992).

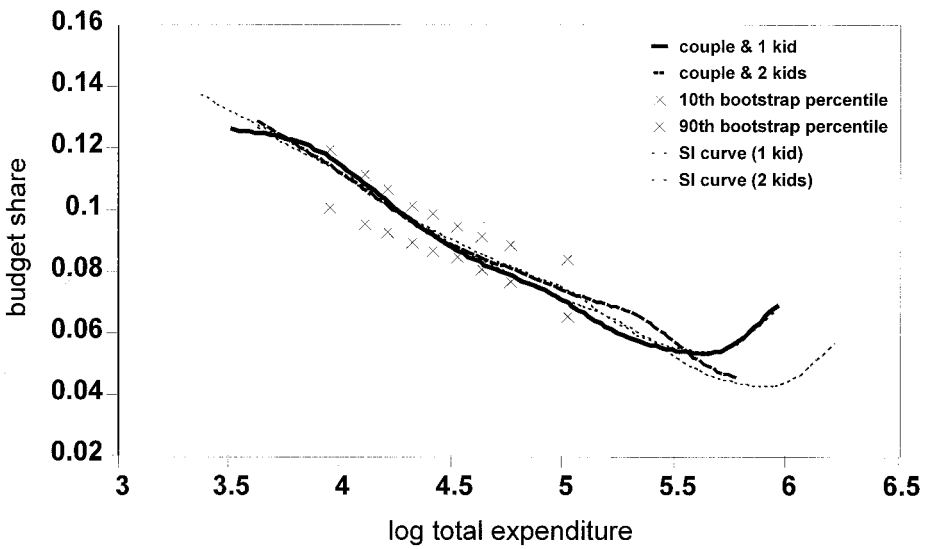

Figure 8. Shape-invariant transformation: fuel share

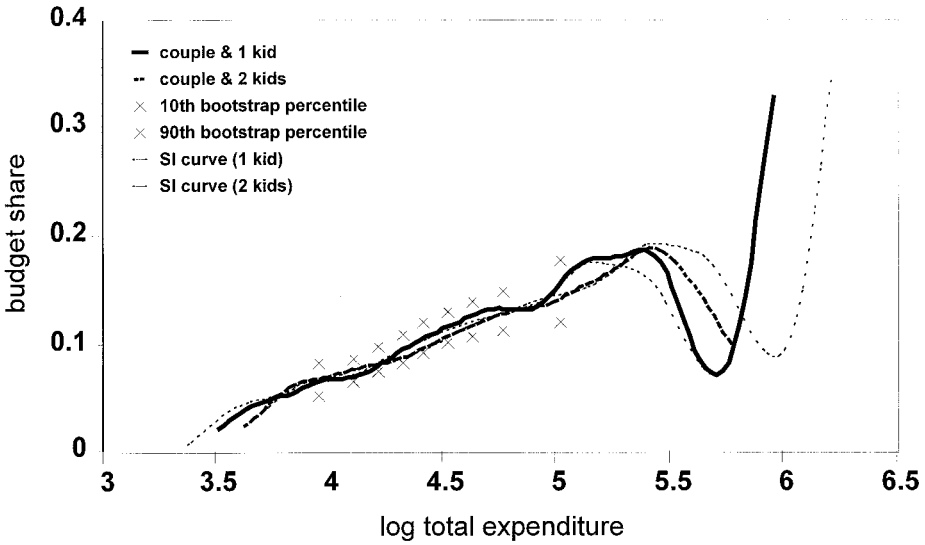

Figure 9. Shape-invariant transformation: clothing share

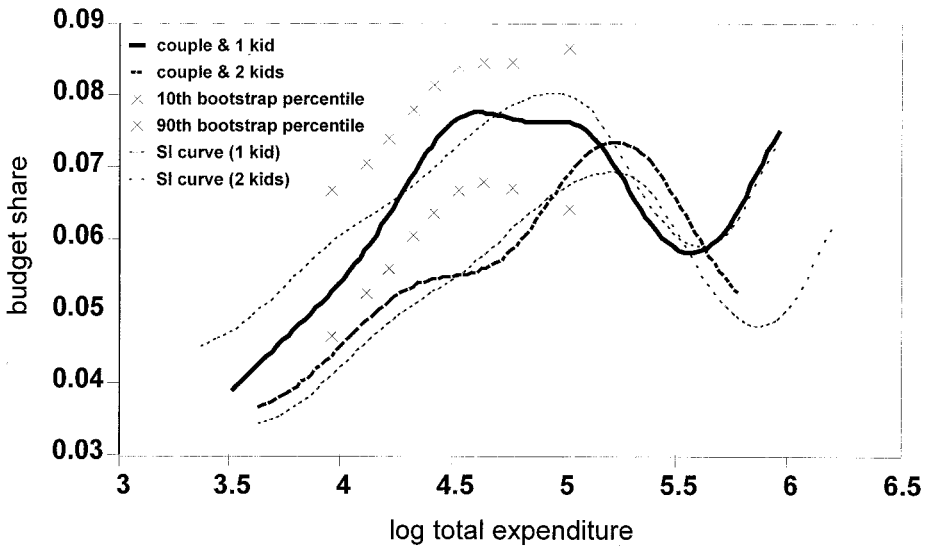

Figure 10. Shape-invariant transformation: alcohol share

expenditure, this being the Engel curve shape underlying the popular AID and Translog demand models of Deaton and Muellbauer (1980a) and Jorgenson, Lau, and Stoker (1980).

We also consider parametric models which have more variety of curvature than is permitted by the Piglog. This reflects growing evidence from a series of empirical studies that suggest quadratic logarithmic income terms are required for certain expenditure share equations. Consequently we

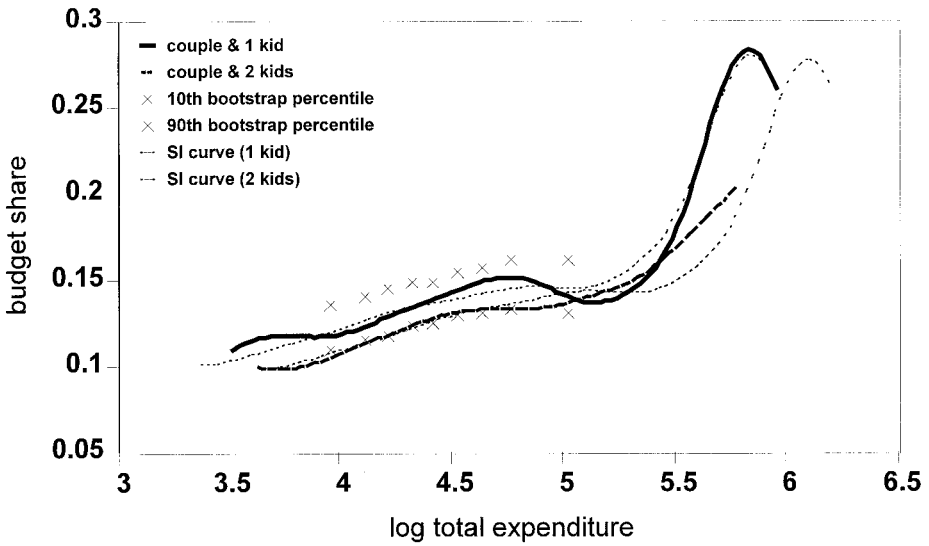

Figure 11. Shape-invariant transformation: transport share

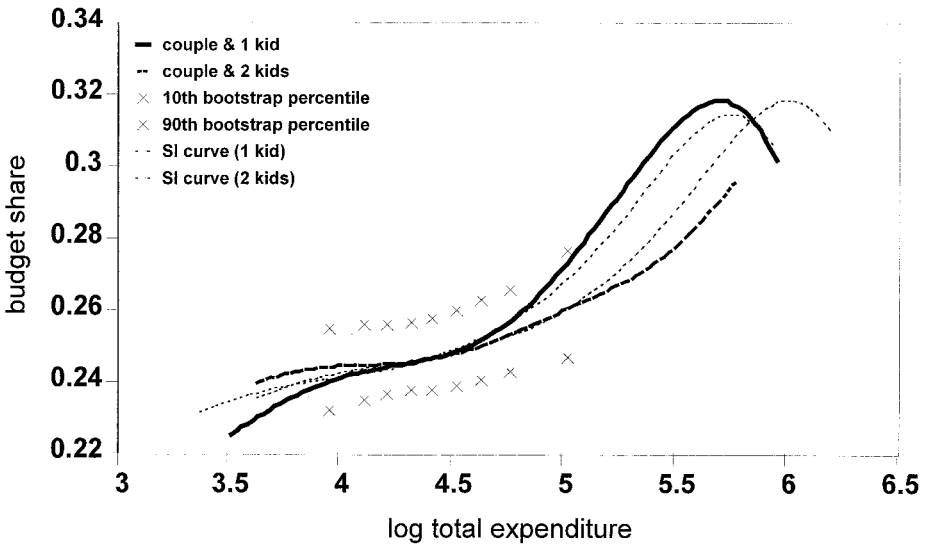

Figure 12. Shape-invariant transformation: other goods share

have used both the Piglog and quadratic logarithmic specifications as null parametric specifications for designing tests against a non-parametric alternative.

Restrictions from consumer theory have been used to place restrictions on the form the Engel curve relationship and the way non-parametric Engel curves can be pooled across demographic types. We have shown that the additive structure between demographic composition and income

that underlies the partially linear semiparametric model implies strong and unreasonable restrictions on behaviour. On the other hand, pooling across demographic types using the shape-invariant semiparametric framework of Härdle and Marron (1990) and Pinkse and Robinson (1995), is shown to provide a preference consistent method for general non-parametric Engel curves. This specification also appears to work well in application.

In the empirical analysis of Engel curves an important issue turned out to be the endogeneity of total expenditure. To account for such endogeneity we have adapted the Holly and Sargan (1982) augmented regression approach to the partially linear regression context. We also considered the Newey, Powell, and Vella (1995) extension to additive recursive structures. In the application, using earned income to instrument total expenditure, correcting for endogeneity is found to have an important impact on the curvature of the Engel curve relationship.

To compare these semiparametric specifications with the Piglog and quadratic logarithmic parametric specifications we implement the recently developed specification test by Aït-Sahalia, Bickel, and Stoker (1994). The Working–Leser or Piglog specification was strongly rejected for some budget shares but the quadratic logarithmic model seemed to provide an acceptable parametric specification.

ACKNOWLEDGEMENTS

Comments from Karim Abadir, James Banks, Ian Crawford, Hide Ichimura, Arthur Lewbel, Joel Horowitz, Joris Pinkse, Arthur Van Soest, three anonymous referees and participants at the Centre workshop are gratefully acknowledged. This work has benefited from the financial support of the ESRC Centre for the Micro-Economic Analysis of Fiscal Policy at IFS. Household data from the FES made available by the CSO through the ESRC Data Archive has been used by permission of the HMSO. Neither the CSO nor the ESRC Data Archive bear responsibility for the analysis or the interpretation of the data reported here. The usual disclaimer applies.

REFERENCES

Aï-Sahalia, Y., P. Bickel, and T. Stoker (1994), 'Goodness-of-fit tests for regression using kernel methods', mimeo, MIT.

Atkinson, A., J. Gomulka and N. Stern (1990), 'Spending on alcohol: evidence from the Family Expenditure Survey 1970–1983', *Economic Journal*, **100**, 808–827.

Banks, J., R. Blundell and A. Lewbel (1997), 'Quadratic Engel curves and consumer demand', *Review of Economics and Statistics*, **79**, 527–539.

Bierens, H. and H. Pott-Buter (1990), 'Specification of household Engel curves by nonparametric regression', *Econometric Reviews*, **9**, 123–184.

Blackorby, C. and D. Donaldson (1993), 'Adult equivalence scales and the economic implementation of interpersonal comparisons of well-being', *Social Choice and Welfare*, **10**, 335–361.

Blackorby, C. and D. Donaldson (1994), 'Measuring the cost of children: a theoretical framework', in R. Blundell, I. Preston and I. Walker (eds), *The Measurement of Household Welfare*, chap. 2, pp. 51–69. Cambridge University Press, Cambridge.

Blundell, R., M. Browning and I. Crawford (1997), 'Nonparametric Engel curves and revealed preference', Discussion Paper W97/14, Institute for Fiscal Studies.

Blundell, R. and A. Duncan (1998), 'Kernel regression in empirical microeconomics', *Journal of Human Resources*, **33**, 62–87.

Blundell, R., P. Pashardes and G. Weber (1993), 'What do we learn about consumer demand patterns from micro data?' *American Economic Review*, **83**, 570–597.

Deaton, A. and J. Muellbauer (1980a), 'An almost ideal demand system', *American Economic Review*, **70**, 312–326.
Deaton, A. and J. Muellbauer (1980b), *Economics and Consumer Behaviour*, Cambridge University Press, Cambridge.
Duncan, A. and A. Jones (1992), 'NP-REG: An interactive package for kernel density estimation and nonparametric regression', Discussion Paper W92/07, Institute for Fiscal Studies.
Ellison, G. and S. F. Ellison (1992), 'A nonparametric residual-based specification test: asymptotic, finite-sample and computational properties', mimeo, Harvard University.
Gorman, W. (1981), 'Some Engel curves', in A. Denton (ed.), *Essays in the Theory and Measurement of Consumer Behaviour*, Cambridge University Press, Cambridge.
Gozalo, P. (1997), 'Nonparametric bootstrap analysis with applications to demographic effects in demand functions', *Journal of Econometrics*, **81**, 357–393.
Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.
Härdle, W. and M. Jerison (1991), 'Cross-sectional Engel curves over time', *Recherches Economiques de Louvain*, **57**, 391–431.
Härdle, W. and E. Mammen (1993), 'Comparing nonparametric vs. parametric regression fits', *Annals of Statistics*, **21**, 1926–1947.
Härdle, W. and J. Marron (1990), 'Semiparametric comparison of regression curves', *Annals of Statistics*, **18**, 63–89.
Hausman, J., W. Newey, H. Ichimura and J. Powell (1991), 'Identification and estimation of polynomial errors in variables models', *Journal of Econometrics*, **50**, 273–296.
Hausman, J., W. Newey and J. Powell (1995), 'Nonlinear errors in variables: estimation of some Engel curves', *Journal of Econometrics*, **65**, 205–234.
Heckman, J., H. Ichimura, J. Smith and P. Todd (1995). 'Nonparametric characterization of selection bias using experimental data: a study of adult males in JTPA', mimeo, University of Chicago.
Holly, A. and J. Sargan (1982), 'Testing for exogeneity in a limited information framework', *Cahiers de Recherches Economiques*, No. 8204, Universite de Lausanne.
Jorgenson, D., L. Lau and T. Stoker (1980), 'Welfare comparison and exact aggregation', *American Economic Review*, **70**, 268–272.
Kneip, A. (1994), 'Nonparametric estimation of common regressors for similar curve data', *Annals of Statistics*, **22**, 1386–1427.
Leser, C. (1963), 'Forms of Engel functions', *Econometrica*, **31**, 694–703.
Lewbel, A. (1989), 'Identification and estimation of equivalence scales under weak separability', *Review of Economic Studies*, **52**, 311–316.
Lewbel, A. (1991), 'The rank of demand systems: theory and nonparametric estimation', *Econometrica*, **59**, 711–730.
Muellbauer, J. (1976), 'Community preferences and the representative consumer', *Econometrica*, **94**, 979–1000.
Newey, W. K., J. L. Powell and F. Vella (1995), 'Nonparametric estimation of triangular simultaneous equations models', mimeo, MIT Department of Economics, forthcoming *Econometrica*.
Pendakur, K. (1998), 'Semiparametric estimates and tests of base-independent equivalence scales', *Journal of Econometrics* (forthcoming).
Pinkse, C. and P. Robinson (1995), 'Pooling nonparametric estimates of regression functions with a similar shape', in G. Maddala, P. Phillips and T. N. Srinivasan (eds), *Advances in Econometrics and Quantitative Economics*, pp. 172–195.
Robinson, P. (1988), 'Root-N-consistent semiparametric regression', *Econometrica*, **56**, 931–954.
Stoker, T. (1991), *Lectures in Semiparametric Econometrics*, CORE Lecture Series. CORE Foundation.
Working, H. (1943), 'Statistical laws of family expenditure', *Journal of the American Statistical Association*, **38**, 43–56.
Zheng, J. (1996), 'A consistent test of functional form via nonparametric estimation techniques', *Journal of Econometrics*, **75**, 263–289.

