Estimation and Testing An Additive Partially Linear Model in a System of Engel Curves.

Jorge Barrientos-Marin*
Dept. Fundamentos del Análisis Económico
University of Alicante, 03080. Alicante, Spain

February 21, 2007

Abstract

The form of the Engel curve has long been a subject of discussion in applied econometrics and until now there has no been definitive conclusion about its form. In this paper an additive partially linear model is used to estimate semiparametrically the effect of total expenditure in the context of the Engel curves. Additionally, we consider the nonparametric inclusion of some regressors which traditionally have a non linear effect such as age and schooling. To that end we compare an additive partially linear model with the fully nonparametric one using recent popular test statistics. We also provide the p-values computed by bootstrap and subsampling schemes for the proposed test statistics. Empirical analysis based on data drawn from the Spanish Expenditure Survey *1990-91* shows that modelling the effects of expenditure, age and schooling on budget share deserves a treatment better than that adopted in simple semiparametric analysis.

Keyword and phrases: Engel curve, expenditure, nonparametric estimation, marginal integration, bootstrap and subsampling.

*I acknowledge financial support from the Spanish Ministry of Education. I am very grateful to Stefan Sperlich for his help and valuable suggestions. I thank LSP members from the University of Paul Sabatier, and especially F. Ferraty and P. Vieu, who provided me with the best atmosphere to prepare this paper. I am solely responsible for the interpretation and for any mistakes. E-mail: jbarr@merlin.fae.ua.es.

1 Introduction

The specification of Engel curves in empirical microeconomics has been an important problem since the early studies of Working (1943) and Leser (1963) and the well-known work of Deaton and Muellbauer (1980a), in which they developed parametric structures such as the Almost Ideal and Translog demand model. Many Microeconomic examples are provided in Deaton and Muellbauer (1980b) in which a separable structure is convenient for analysis and important for interpretability. However, there is increasing empirical evidence pointing to the conclusion that a sort of nonlinearity is present in the specification of Engel curves. An alternative way of investigating nonlinear effects is to model consumer behavior by means of semi- and nonparametric additive structures. Moreover, non and semiparametric regression provides an alternative to standard parametric regression, allowing the data to determine the local shape of the conditional mean.

From an economic point of view there are many reasons why it is interesting to recover a correct specification of Engel curves. Firstly, a correct specification allows us to examine the nature of the effect of changes in indirect tax reforms. Secondly, it is important to specify the response of consumers in the face of changes in total income. Changes of this kind allow us to assess the impact on consumers' welfare.

Consumer demand has become a very important field for applying non and semiparametric methods. An interesting analysis of the cross-sectional behavior of consumers in the context of a fully nonparametric model can be found in Bierens and Pott-Buter (1990). Papers which consider the implementation of semiparametric methods in empirical analysis of consumer demand include Banks, Blundell and Lewbel (1997) and Blundell, Duncan and Pendakur (1998). This latter paper is of special interest because its analysis regression is based on semi- and nonparametric specifications of Engel curves. It also tests Working-Leser and Piglog's null hypothesis against the well-known partial linear model in which budget expenditures are linear in the log of total expenditure. In this paper we estimate the Engel curves directly as in Lyssiotou, Pashardes and Stengos (2001) among others.

We estimate an additive partially linear model (PLM) in order to investigate consumer behavior using individual household data drawn from the Spanish Expenditure Survey (SES) and use the result obtained from semiparametric analysis to examine the modelling of age, schooling and expenditure in a system of Engel curves. The importance of using an additive

PLM models lies in the fact that in the context of this model the effects of expenditure, the age and schooling on consumer demand can be investigated simultaneously in the semiparametric context[1]. There are several ways to get estimations of nonparametric additive structure, and we mention only the most important: smooth backfitting, series estimators and marginal integration. In this paper we use internalized marginal integration to estimate nonparametric components in the additive PLM mainly because at the present time there is no applied or theoretical study on the testing procedure using smooth backfitting.

Most of the papers that investigate consumer behavior in a nonparametric context are focused on the appropriate way of modeling the form of the Engel curves. Those focused on the unidimensional nonparametric effect of log total expenditure on budget expenditures, taking in to account some parametric indexes to reflect demographic composition include Blundell, Browning and Crawford (2003) and references therein. In this paper we investigate consumer behavior in semi and -nonparametric terms focused on the nonparametric effect of total expenditure the age and the schooling. In this study, unless stated otherwise, the effect of age and schooling refer to the age and schooling of the household head. There is evidence suggesting that these have deeper effect than generally assumed in parametric demand analysis (see Lyssiotou, Pashardes and Stengos (2001)). In fact, it is common practice to include the square of age and/or schooling as well as their higher terms in parametric models to capture possible nonlinear effects.

Inference in nonparametric regression can take place in a number of ways. The most natural is to use nonparametric regression as an alternative against a fully parametric or semiparametric null hypothesis. With this in mind, we investigate whether an additive PLM provides a reasonable adjustment to our data using different resampling schemes to obtain critical values of the test statistics. In this paper we are interested in applying some recently developed test statistics which are very popular in the literature about testing semiparametric hypotheses against nonparametric alternatives. These test statistics are in the spirit of Hardle and Mammen (1993) and Gozalo and Linton (2001), among others. On the other hand there is a growing interest in the so called adaptive testing methods, in which the test statistics

[1]Analysis of consumer behavior can be carried out with fully nonparametric models. However, for sake of interpretability and implementation, additive models overcome the well-known problems coming from multidimensional Nadaraya-Watson and Local Polynomial regression estimators.

are adaptive to the unknown smoothness of the alternative, see among others Horowitz and Sponkoiny (2001) and Rodriguez-Poo, Sperlich and Vieu (2005). In this paper we adapt their ideas with some differences, where are considered kernel smoother for our problem.

It should be remarked that a problem that we may well have to consider is the endogeneity of regressors. Note that in the context of Engel curves total expenditure may well be jointly determined with expenditure on different goods. The approach used to solve this problem is instrumental variable estimation. We remark two recently developed procedures in the context of nonparametric regression to tackle the problem of endogenous regressors. The so called nonparametric two step least square (NP2SLS) due to Newey and Powell (2003), and the nonparametric two step with generated regressors and constructed variables (NP2SCV) due to Sperlich (2005). Newey's approach is a cumbersome procedure involving the choice of basis expansion in the first step. However, Sperlich's approach only requires a non, semi or even parametric construction of regressors of interest in the first step. Our feeling is that a generated variables approach in combination with additive PLM can help us to overcome to some extent any possible endogeneity problem and that is exactly the procedure implemented in this paper.

The contribution of this work can be summarized as follows. Firstly, we are the first (to our knowledge) to carry out an exploratory analysis of consumer behavior with data drawn from the Family Expenditure Survey for Spain using semiparametric models. Second, we apply recently developed methods to estimate, test (various model specifications) and correct for possible endogeneity of total expenditure. Third, our estimations of the additive model are accompanied by a reasonable measurement of discrepancy between the fully nonparametric model and the additive estimation. An adequate model check is necessary whenever estimations of additive models are carried out (Dette, von Lieres and Sperlich (2004)). Additionally, our measure of discrepancy adapts to the unknown smoothness of the non-parametric model and this constitutes a novelty in empirical economics.

The rest of the paper is organized as follows. In Section 2 we provide some background to understand both the estimating and the testing procedures. In Section 3, we discuss the shape of Engel curves and report empirical results obtained from the application of additive PLM. We also provide the results of testing the additive specification as well as the linearity of each nonparametric component in additive PLM regression. In Section 4 concludes.

2 Additive Partially Linear Model and Testing Hypothesis

There are many fields of empirical economics in which explanatory variables and their second power are included in regression analysis to capture nonlinear effects; Economics of Education, Return on Education, Labor Economics, and more examples can be given. In these particular examples regressors such as age, schooling or experience (generally measured in years) enter into the linear specification in quadratic form (or in polynomial form with higher terms). The additive model has a structure that is appropriated for capturing the effect of these regressors nonparametrically (not necessarily linearly). Consider the following model:

$$Y_i = m\left(X_i\right) + u_i \qquad i = 1, 2, ..., n, \qquad [1]$$

where $\{Y_i\} \in \mathbb{R}$ is a scalar response, $\{X_i\} \in \mathbb{R}^d$ is a sequence of random variables, $m : \mathbb{R}^d \longrightarrow \mathbb{R}$ is an unknown function and $\{u_i\}$ is an unobserved independent random variable with mean zero. Let ψ be a parameter and $m\left(x, \psi\right)$ an unknown function denoting a semiparametric structure. For the sake of notation we establish that $m\left(x, \psi\right) = m_S(x)$. Then $m\left(x\right)$ has an additive structure if:

$$m_S\left(x\right) = E\left(Y|X = x\right) = \psi + \sum_{\alpha=1}^{d} m_\alpha\left(x_\alpha\right) \qquad [2]$$

The structure of the model in eq.[2] was first discussed by Stone (1985, 1986) who shown that the additive components can be consistently estimated at the same rate as in a one dimensional fully nonparametric regression model. Linton and Nielsen (1995) propose estimating the additive components of the eq.[2], in a bidimensional context, by marginal integrating a local estimator of $m\left(\cdot\right)$. In general terms the integration idea is based on the following observation. Let $X = \left(X_1, ..., X_d\right)^T$ be a vector of explanatory variables, $\left\{m_\alpha\left(\cdot\right)\right\}_{\alpha=1}^{d}$ a set of unknown function satisfying $E_{X_\alpha}\left\{m_\alpha(X_\alpha)\right\} = \int m_\alpha(x_\alpha) f_\alpha(x_\alpha) dx_\alpha = 0 \ \forall \alpha \in \Lambda$ and $E\left\{Y\right\} = E\left\{m_\beta(X_\beta)\right\} = \psi$ for identification. Then, if $E\left(Y|X = x\right)$ is additive and the marginal density of X_β is

denoted by $f_\beta(\cdot)$, for a fixed x point we have that:

$$\begin{aligned} E_{X_\beta}\{m(X_\alpha, X_\beta)\} &= \int m(x_\alpha, x_\beta) f_\beta(x_\beta) \prod_{\beta \neq \alpha} dx_\beta \\ &= \psi + m_\alpha(x_\alpha) + \sum_{\beta \neq \alpha} 0 \end{aligned} \qquad [3]$$

In order to estimate the functions $m_\alpha(x_\alpha)$ we first estimate the function $m(x)$ with a multidimensional local smoother and then integrate out the variables different from X_α. This method can be applied to estimate all the components, and finally the regression function $m(\cdot)$ is estimated by summing an estimator $\hat{\psi}$ of ψ, so we get that:

$$\hat{m}_S(X_j) = \hat{\psi} + \sum_{\alpha=1}^{d} \sum_{i=1}^{n} K_h(X_{j\alpha} - X_{i\alpha}) \frac{\hat{f}_\beta(X_{i,\beta})}{\hat{f}(X_{i\alpha}, X_{i,\beta})} Y_i \qquad [4]$$

for *j=1,...,n*. The expression to get the estimation of each component $m_\alpha(\cdot)$ defined in [4], is called the internalized marginal integration estimator (IMIE) because of the joint density that appears under the summation sign. For a detailed explanation see Dette, von Lieres and Sperlich (2004) and references therein. Note that IMIE does not provide exactly the orthogonal projection onto the space of additive functions. In other words, the sum of the estimated nonparametric components does not necessarily recover the complete conditional mean because the interaction terms are excluded from the regression. So, it is very interesting to establish whether the sum of additive components is the conditional mean. Therefore, it is necessary to carry out a specification test. With this in mind, we are concerned with testing the validity of the additive specification of the regression function $m(x)$ in eq.[1]. Thus, the null hypothesis to be tested can be formulated as:

$$H_0 : m(x) = m_S(x) \qquad [5]$$

against a general alternative that H_0 is false. An adaptive test statistics is implemented by Horowitz and Spokoiny (2001) (among others) in the context of parametric models against nonparametric alternatives; and by Rodriguez-Poo, Sperlich and Vieu (2005) in the context of semi and nonparametric against a nonparametric alternative. However, it should be remarked that

the first implementation in the context of nonparametric additive separable models against a fully nonparametric alternative adaptive test was by Barrientos and Sperlich (2005).

The first test statistic is defined as the square of the differences between the semiparametric fit and the fully nonparametric estimator, extending the concept introduced by Hardle and Mammen (1993). In order to test the validity of our hypothesis, we also consider the test statistics introduced by Gozalo and Linton (2001) and Rodriguez-Poo, Sperlich and Vieu (2005) defined as:

$$\hat{T}_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}(X_i) - \hat{m}_S(X_i)\right]^2 w(X_i) \qquad [6]$$

$$\hat{T}_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i\left[\hat{m}(X_i) - \hat{m}_S(X_i)\right] w\left(X_i\right) \qquad [7]$$

$$\hat{T}_3 = \sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}K_h\left(X_i - X_j\right)\left(Y_j - \hat{m}_S\left(X_j\right)\right)\right]^2 w\left(X_i\right) \qquad [8]$$

where $\hat{e}_i = Y_i - \hat{m}_S(X_i)$ are the residuals under the additive model and $\hat{u}_i = Y_i - \hat{m}_k^I(X_i)$ denote the corresponding residuals of the unrestricted model. In this study we use the well-known Nadaraya (1964)-Watson (1964) estimator for the unrestricted model. These test statistics can be used not only in specification testing defined by [5] but also to test the linearity of individual nonparametric components, see Hardle, Huet Mammen and Sperlich (2004). More exactly, we can test the null hypothesis

$$H_0 : m_\alpha\left(x_\alpha\right) = \theta x_\alpha \text{ for all } \alpha \text{ and for some } \theta$$

Now we discuss the procedure for computing the critical values. Note that our idea is based on a combination of adaptive test statistics with both

bootstrap[2] and subsampling[3] schemes. For the former case see Horowitz and Spokoiny (2001) and for the latter one see Rodriguez-Poo, Sperlich and Vieu (2005). It is remarkable that using subsampling to get an estimator of the variance of the restricted errors guarantees consistency under H_1. Having estimated semiparametric and nonparametric models, $\hat{m}_S(\cdot)$ and $\hat{m}(\cdot)$ respectively, we construct the original test statistics denoted by $\hat{T}_{jk}$. As the distribution of $\hat{T}_{jk}$ varies with k we define the standard test statistic denoted by

$$\hat{\tau}_{jk} = \frac{\hat{T}_{jk} - \hat{\mu}_j}{\hat{v}_j} \qquad [9]$$

where $\hat{\mu}_j$ and $\hat{v}_j^2$ are the estimated mean and variance of the test $\hat{T}_{jk}$ for $j = 1, 2, 3,$. Then we compute the test statistics based on the resampling data (boootstrap and subsampling data), denoted by:

$$\hat{\tau}^*_{jk} = \frac{\hat{T}^*_{jk} - \hat{\mu}^*_j}{\hat{v}^*_j} \qquad [10]$$

This creates a family of test statistics $\{\tau_k,\, k \in K_n\}$ where the choice of k makes the difference between the null and global alternative hypotheses. In order to maximize power we take the maximum of $\hat{\tau}^*_{jk}$ over a finite set of bandwidth values K_n with cardinality L. Then we define the final test statistics by means of:

[2]To obtain bootstrap critical values we consider the following steps. 1) To obtain the bandwidth from cross-validation, h_{cv}. 2) To estimate $\hat{m}_S(x) = \hat{\psi} + \sum_{\alpha \in \Lambda} \hat{m}_\alpha(x_\alpha)$ .3) To use the bootstrap scheme to get ε^*_i for each $i = 1, ..., n$. 4) For each $i = 1, ...n$ generate $Y^*_i = \hat{m}_S(X_i) + \varepsilon^*_i$, where ε^*_i is sampled randomly and we use the data $\{Y^*_i, X_i\}_{i=1}^n$ to estimate $\hat{m}_S(x)$ under H_0. 5) Repeat the process 2-4 B times to obtain $\{\tau^*_{jk}\}$ and use these B values to construct the empirical bootstrap distribution.

The bootstrap errors ε^*_i are generated by multiplying the original estimated residuals from the semiparametric model, $\hat{\varepsilon}_i = Y_i - \hat{m}_S(X_i)$, by a random variable with standard distribution. This procedure provides exactly the same first and second moments for $\hat{\varepsilon}_i$ and for ε^*_i.

[3]In the subsampling case one takes all subsamples of size b from the original sample $\{X_i, Y_i\}$. The problem in selecting the subsample size b is similar to the problem in selecting the bandwidth in nonparametric regression analysis: the assumptions on the parameter b to require that $b/n \longrightarrow 0$ and $b \longrightarrow \infty$ as $n \longrightarrow \infty$. Unfortunately, such asymptotic conditions are no help in solving the block size choice problem in finite samples. Instead, it is possible to use an algorithm to estimate a "good" subsample size. This method has been applied in practice in another contexts with good results, see for instance Rodriguez-Poo, Sperlich and Vieu (2004) and Neumeyer and Sperlich (2005).

$$\hat{\tau}_{jk}^{**} = \max_{k \in K_n} \hat{\tau}_{jk}^{*} \qquad [11]$$

Where $K_n = \left\{k = a_{(l)} n^{-1/5} \quad l = 1, ..., L\right\}$, $a(l) = \left[l + \left(c_X (l-1)^{-1}\right)\right] n^{-1/5}$ and $c_X = \gamma (\max(X_i) - \min(X_i))$ with $\gamma \in (0, 1)$.

The testing procedure rejects H_0 if at least one of the $k \in K_n$ the original test statistic is significantly larger than the bootstrap analogues. In Horowitz and Spokoiny (2001) the estimators for variance and bias are asked to be consistent under alternative hypothesis. Note that this is only necessary for efficiency; for consistency of the test, it is sufficient for the difference between real variance and estimate to be bounded. Nevertheless, Rodriguez -Poo, Sperlich and Vieu (2005) suggest using a subsampling scheme in order to get a consistent estimator of variance under H_1 and thus to have optimal power. They also discuss size problems of bootstrap tests when the null model is non or semiparametric and show that the subsampling based analogue suffers less from this problem.

3 The Shape of Engel Curves and Specification Testing

The most usual structure in consumer behavior analysis is the so-called Working-Leser specification. In this model each expenditure expenditure is defined over the logarithm of total expenditure. Thus the model has a simple structure given by:

$$w_i = f(\ln X_i) + \varepsilon_j \qquad [12]$$

where w_i is the budget expenditure, $\ln X_i$ is the log total expenditure and ε_i is an error term satisfying $E(\varepsilon_i | \ln X_i) = 0$. Empirical analysis using parametric specification in eq.[12] can be found in the literature on consumer behavior, see Deaton and Muellbauer (1980a, 1980b). For empirical unidimensional nonparametric analysis see Blundell, Browning and Crawford (2003) and references therein. Instead of a Working-Leser specification we can assume that consumer demand could be modelled by means of an additive structure as in eq[2], such that:

$$w_i = \psi + m_1(\ln X_{1i}) + m_2(X_{2i}) + m_3(X_{3i}) + \varepsilon_i \qquad i = 1, ..., n \qquad [13]$$

where $\ln X_{1i}$ is the log total expenditure, X_{2i} and X_{3i} are the age and schooling and ε_i is assumed to satisfy $E(\varepsilon_i|X_i) = 0$. Consider the augmented model:

$$w_i = \psi + Z_i\beta_k + m_1(\ln X_{1i}) + m_2(X_{2i}) + m_3(X_{3i}) + \varepsilon_i \qquad [14]$$

$i = 1, ..., n$ where Z_i is a set of discrete or continuous variables of dimension K, β is a $K \times 1$ vector of parameters and ε_i is assumed to satisfy $E(\varepsilon_i|Z_i,\ X_i) = 0$. The models given by [13] and [14] are motivated because they allow us to include other regressors with nonlinear effects, and at the same time to reduce the curse of dimensionality; which may be the main weakness of nonparametric techniques. To estimate the model [14] we follow the treatment of Hengartner and Sperlich (2005). There are many ways to get a $\sqrt{n}$-consistent estimator of β: we use Robinson's (1988) method. Let $\hat{\beta}$ be an estimator of β. Eq.[14] can be written as:

$$\omega_i = \psi + m_1(\ln X_{1i}) + m_2(X_{2i}) + m_3(X_{3i}) + \xi_i \qquad [15]$$

where $\omega_i = w_i - Z_i\hat{\beta}_k$ and $\xi_i = \varepsilon_i + Z_i\left(\hat{\beta}_k - \beta_k\right)$ is the new composite error term. The intercept term ψ can be $\sqrt{n}$-consistently estimated by $\hat{\psi} = \bar{Y} - \bar{Z}^T\hat{\beta}$ where $\bar{Y}$ and $\bar{X}$ are the sample mean. As in eq.[14] we can apply to eq.[15] the procedure described in Section 2 to obtain estimates of $m_1(\ln X_1)$, $m_2(X_2)$ and $m_3(X_3)$.

Now we turn to the problem mentioned in the Introduction about constructing regressors to overcome the endogeneity problem. For a detailed explanation, see Sperlich (2005). Let x_i be an unobservable or endogenous variable and let $\hat{X}_i$ be a generated regressor[4], it is then possible to write $\hat{x}_i = x + b(x) + \sigma(x)$, where $b(\cdot)$ is the bias term such that $b(\cdot) \to 0$ as $n \to \infty$ and $\sigma(\cdot)$ is the variance term. In order to obtain consistent estimates of density and conditional mean and thus construct $\hat{x}_i$, with the help of instruments or even with help from different data sets it is possible to estimate the reduced regression form, semi-, non- or even parametrically (first step) and then use it in the structural regression (second step), instead of the original regressor.

The procedure can be described as follows. Let $\{W\}$ be the set of exogenous variables, including the log of total income. Note that we are worried

[4] with bias and variance of order $O(g^2)$ and $O\left(\frac{1}{ng^\delta}\right)$ in order to fulfil the assumptions of Theorem 2 in Sperlich (2005)

about endogeneity due to jointly determination of total expenditure, $\ln X_1$, and expenditure on different categories of goods (endogeneity due to simultaneity). Suppose that $\ln X_1$ is endogenous such that:

$$\ln(X_1) = g(W) + U \qquad [16]$$

where $E(\xi|W) = E(U|W) = 0$ for $W = (Z, X_2, X_3)$, but $E(\xi|\ln X_1) \neq 0$. Putting [15] and [16] together we get

$$\omega = \psi + m_1(g(W) + U) + m_2(X_2) + m_3(X_3) + \xi$$

applying the modeling $m_1(g(W) + U) = m_g(g(W)) + \lambda(U)$, what somehow means in the end that you assume additive in the exogenous impact of the explanatory variables what is possible as assuming [16]. The we get:

$$\omega = \psi + m_g(g(W)) + m_2(X_2) + m_3(X_3) + \lambda(U) + \xi \qquad [17]$$
$$\omega = \psi + m_g(g(W)) + m_2(X_2) + m_3(X_3) + \tilde{\xi} \qquad [18]$$

where $E[\lambda(U)] = 0$ and $E[\xi|g(W), X_2, X_3] = 0$. The expression in eq[18] is the model that we have estimated only with the additional burden of a pre-estimation $g(W)$ consistently. Note that this methodology certainly involve less difficulties (and is faster) than Newey and Powell's (2003) approach.

Household expenditures typically display variation respect to demographic composition. Then, we can use additive specification to pool across household types. However, Blundell et.al (2003) suggested modifications to take into account integrability conditions (integrability is related to the problem of recovering a consumer's utility function from his demand functions). Note that in eq[14], the Z matrix represent a household composition variables for each household observation i. This means that we imposed a restriction on the way in which demographics affect expenditures (if j index is referred to specific category of good then we are interested in imposing the restriction $Z_i = Z_{ij}$, that is demographic composition affects in the same way the consumption of different goods). Thus, under stated restriction on Z matrix, our empirical researching did not provide evidence of linearity of $m_1(\cdot)$ in our system (see Section 3 and Table 4).

Blundell et al. (2003) agrees that an alternative specification that does not impose restriction on the form of $m_1(\ln X_1)$ is a straightforward extension of additive PLM: $w_i = \psi + Z_i\alpha_k + m_1(\ln X_{1i} - \xi(Z_i'\theta)) + m_2(X_{2i}) + m_3(X_{3i})$

in which $\xi(Z_i'\theta)$ is some known function[5] of a finite set of parameters θ (otherwise $m_1(\cdot)$ might be linear in $\ln X_1$ whenever Slutsky symmetry conditions are satisfied).

3.1 Data Used in this Application

In our application we consider mainly four broad categories of goods, Food (including alcohol and tobacco), Clothing (including shoes), Transport (personal and public) and Leisure (recreational activities, publications and general teaching). We draw data from the 1990-1991 Spanish Expenditure Survey (SES) and for the purposes of our study we select only houses with three children or less. Total income, total expenditure and expenditure categories are measured in pesetas (yearly) at constant 1983 prices. In order to preserve a degree of homogeneity in most of aspects we use a subset of married (or cohabiting) couples of household in the Madrid regional community. This leaves us with 757 observations, 12.4% comprising couples without children, 20.02% couples with one child, 47% couples with two children and 20.03% couples with three children. Table 1 gives brief descriptive statistics for the main variables used in the empirical analysis.

[5] As they suggested $\xi(Z_i'\theta)$ can be interpreted as the log of a general equivalence scale for household i.

Table 1. Descriptive statistics for budget expenditure data				
Variables	Mean	Std.dev	Min	Max
Food expenditure	709216	348565	344776	3307304
Clothing expenditure	304160	335535	7200	2254260
Transport expenditure	413226	486898	3640	2426126
Leisure expenditure	231988	265513	999	2128000
Total Expenditure	3162401	1397284	1039319	9304396
Log total Expenditure	14.87	0.429	13.85	16.04
Total income	2052240	2289599	282504	42000000
Log total income	14.37	0.50	12.5	17.5
HHAge	40.6	10.6	21	80
HH Schooling	5.1	2.47	1	10
HNAD	2.1	0.5	1	4
HHSEX	0.90	—	0	1
Child_0	0.124	—	0	1
Child_1	0.202	—	0	1
Child_2	0.470	—	0	1
Child_3	0.203	—	0	1

3.2 Some Pictures of the Expenditure expenditure-Log Total Expenditure Relationship

In this section we present the estimated additive partially linear regression of the Engel curves for the four budget expenditures in our SES sample. Each figure presents the estimated marginal effect together with 90% bootstrap pointwise confidence bands (dashed lines). In all cases we present kernel regression for the quartic kernel $\frac{15}{16}\left(1-u^2\right)^2 I\left(|u| \leq 1\right)$ where $I\left(\cdot\right)$ is the indicator function, using the leave-one-out cross-validation method to automatic bandwidth choice, $h_{cv} = 0.72$ in the direction of interest and $b = 6h_{cv}$ in the nuisance direction as in Dette von Lieres and Sperlich (2004). In order to estimate the parametric part of the model [13] we have used a set of discrete variables such as number of adults, sex and dummies for number of children; this kind of regressor traditionally enters into the regression function in the parametric part.

As usually, it is assumed that income is partially correlated with expen-

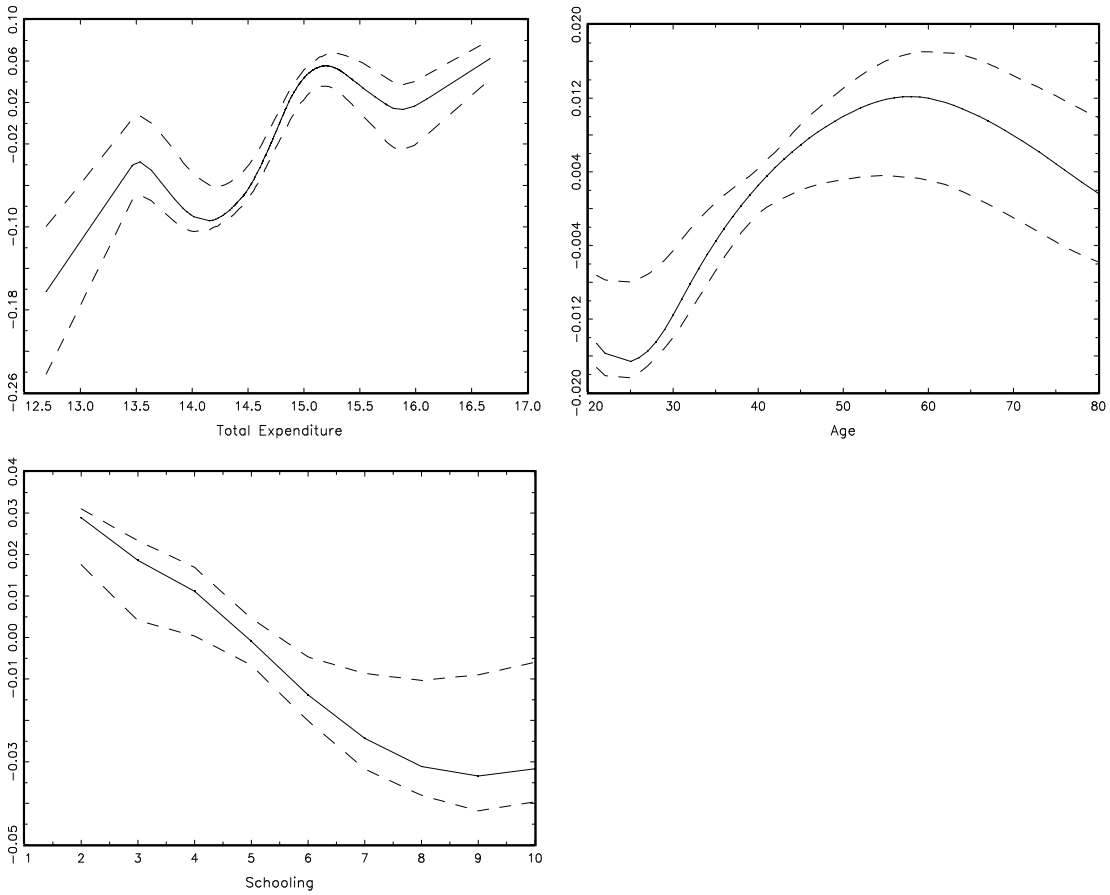

Figure 1: Estimated marginal effect of total expenditure, age and schooling on food expenditure

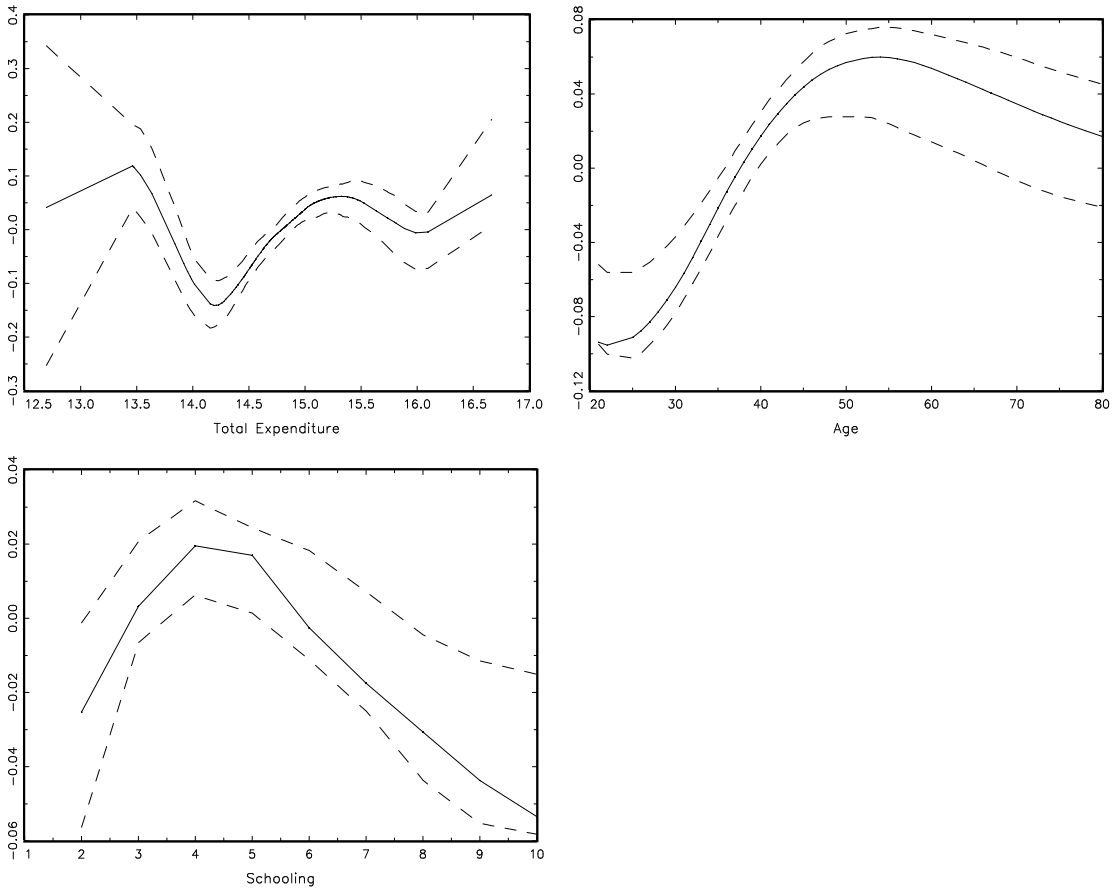

Figure 2: Estimated marginal effect of total expenditure, age and schooling on clothing expenditure

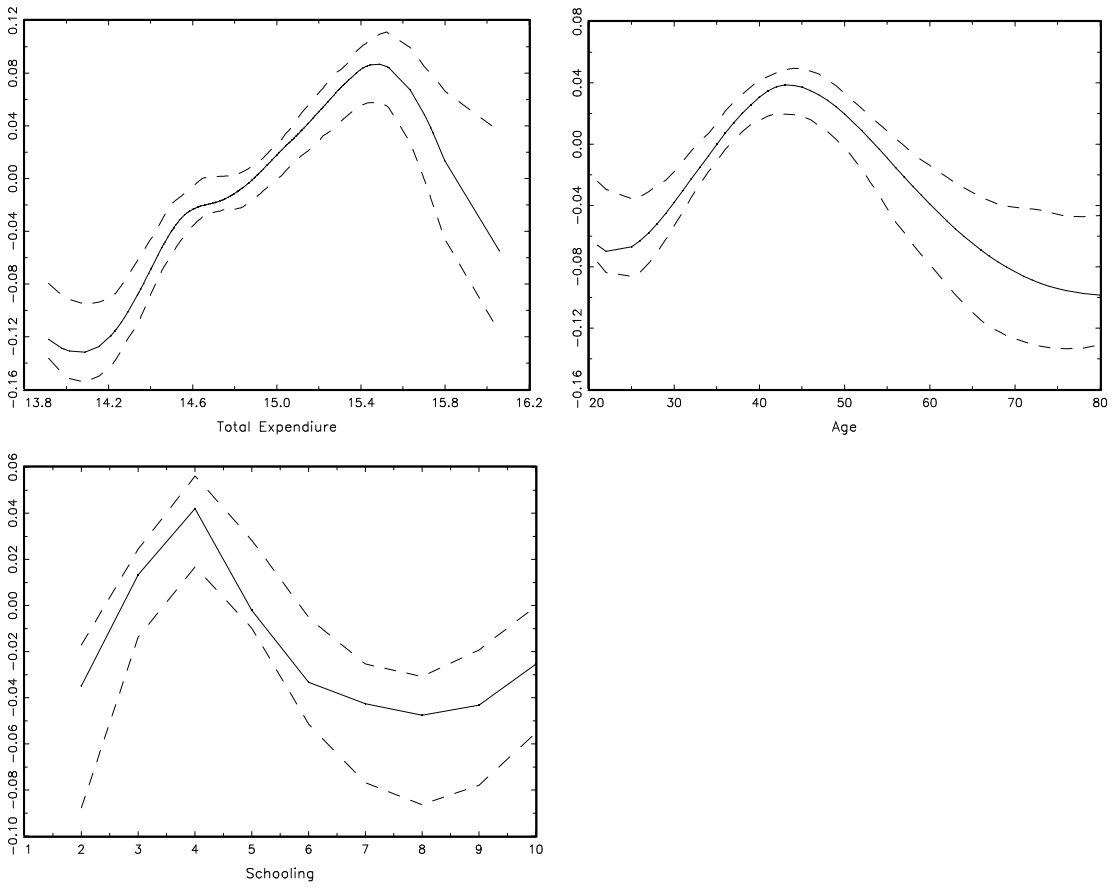

Figure 3: Estimated marginal effect of total expenditure, age and schooling on leisure expenditure

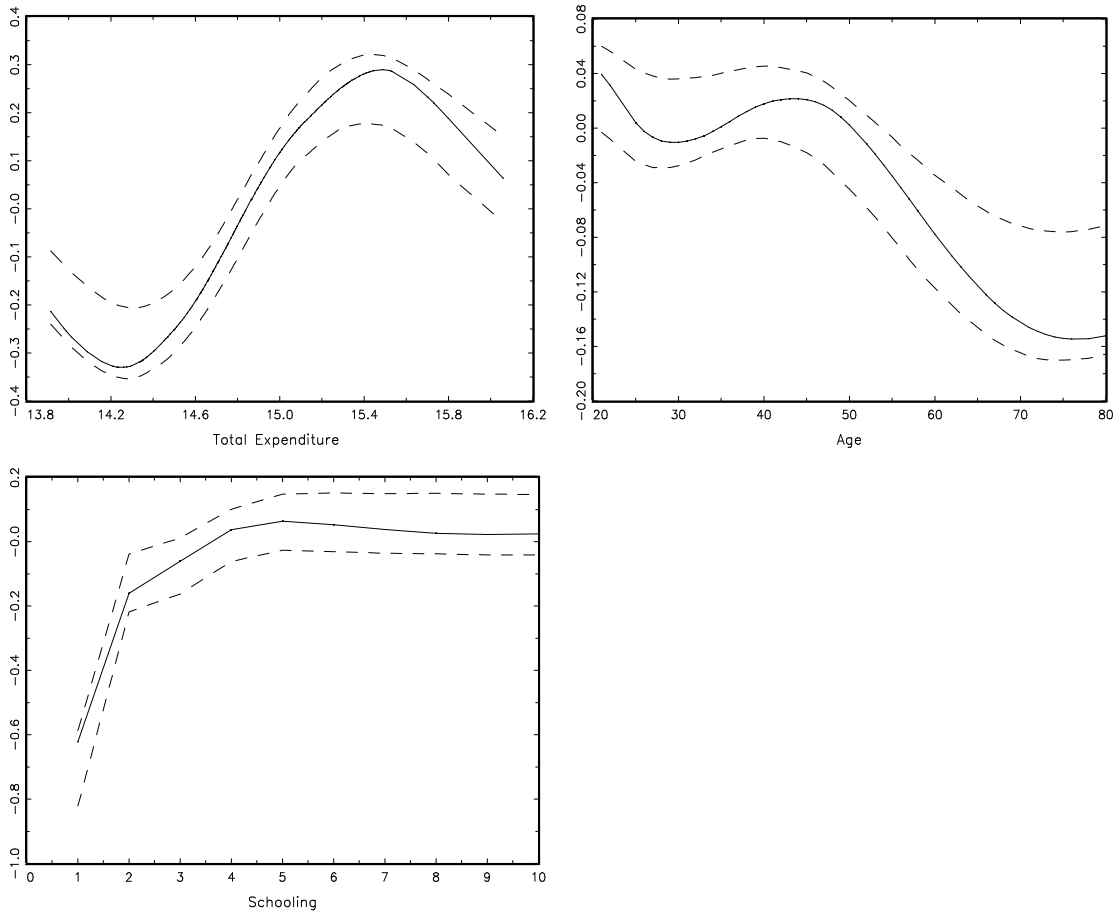

Figure 4: Estimated marginal effect of total expenditure, age and schooling on transport expenditure.

diture and we can suppose that it is not correlated with the errors in model [13], therefore log total income is a natural instrument to the log total expenditure. Then, based on generated regressor and constructed variable methods we adjust the estimations for any possible endogeneity of log total expenditure with the existing data as described in Section 3.1. The set of exogenous variables includes the log income and its power (up to the fourth one), age and schooling[6].

Figures 1-4 show the estimated marginal effect of log total expenditure, age and schooling on budget expenditures, controlled parametrically by the sex, number of adults in households and dummies for number of children and corrected for any possible endogeneity. It is clear from the plots that the effect of total expenditure on the different budget expenditures is nonlinear. We can see in the case of transport and leisure expenditures that this effect is increasing and monotone, whereas in the case of food and clothing it is also increasing, but less stable for different levels of total expenditure.

Note that the effect of schooling on expenditure on different goods is nonlinear. In the cases of leisure, clothing and even transport it is interesting to observe the pronounced effect for values of schooling close to the average (at which point the greatest expenditure expenditure is reached). Note that leisure and clothing are necessary goods (but not basics like food), so this behavior could be related to low returns on education (whenever there is a strong correlation between income and schooling, such a relation is generally observed in practice), so that consumers prefer to dedicate their budget to basic goods. We remark that food expenditure does not include food outside the household. It might be assumed that head of the household might take some meals (e.g lunch) outside the house. On the other hand, in the case of transport expenditure, we note an increasing effect up to values close to the average for schooling, but from that point onwards the expenditure becomes stabilized.

Another possible explanation for the behavior of leisure expenditure with respect to schooling, is that high levels of schooling in couples that have many children are accompanied by high income levels, and more hours of work per week, so that they have no time for leisure. This idea is not so absurd if we consider that more than half of households (67.03%) have two or three children to support.

[6]Estimations for the model given by [13] with no endogeneity correction are available from the author on request.

According to our results, in the households with the oldest heads there is a tendency to spend less money than in the households with younger heads, this effect is notable at least for a range of ages between 30 and 40. It is explained, at least in part, because the households with the oldest heads have less children to support. Unlike leisure and transport expenditures, in the cases of food and clothing expenditures this decreasing effect is considerable but not dramatic. Note that we include the number of children parametrically, so this explanation makes sense if we keep other effects unchanged. However, except for food and clothing expenditures, the estimated parameters have no major impact.

Another question to take in to account is that 90% of household heads in our SES sample are men: from the sociological point of view they pay less attention to fashion, so this may explain, partly, the decrease in spending on clothes for household heads of 40 and over. In the case of leisure and transport, the effect by ages is dramatically decreasing: of course older heads have less recreational activities and spend less time outside the household, so the use of transportation (private and public) diminishes with age.

In regard to variables included parametrically, we remark that the number of adults has no effect on consumer demand; the estimated parameter in each regression is not statistically significant. On the other hand, the effect of sex is important and different depending on the expenditure considered (except for transport). The results tell us that men spend less money on food than on clothing and leisure.

If the model is chosen correctly, the results quantify the extent to which each variable affects consumer behavior. Clearly, the findings of the estimated additive PLM have to be checked: this can be done by considering the test statistics described in Section 2.

3.3 Specification Testing

Table 3 reports the p-values for testing additivity adjusted for any possible endogeneity problem. Since the choice of bandwidth is a crucial point, especially for the bootstrap needed for the test procedure, we present results for different smoothing parameters $g_r \in \{0.75, 0.85, 0.95\}$ *r=1,2,3*. In order to apply the procedure described in Section 2 we implement 500 bootstrap replications; and we use 100 subsamples each of 70% and 60% of the size of the original sample n for our subsampling scheme. To estimate the model under alternative hypotheses (fully nonparametric model) we define a set K_n

(with cardinality L=*10*) of bandwidths k in a range from *0.3* to *2*.

Note that the percentage of rejection is not so large for leisure and transport expenditures with the bootstrap scheme. However, this situation is partly corrected with the subsampling scheme where the percentage of rejection is increases, especially in the case of leisure. On the other hand, we find that test statistics τ_1 and τ_3 give us a strong evidence of additive separable specification. Similar results are obtained with the subsampling scheme in the sense that we are able to reject the null hypothesis for all test statistics for each expenditure categories. In summary, the null hypothesis is not rejected for the household types considered for all test statistics with both resampling schemes and for all bandwidths.

Table 3. Testing Additive Specification

Bootstrap												
Band	Food			Clothing			Leisure			Transport		
	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3
g_1	.66	.81	.99	.95	.22	.99	.92	.13	.99	.97	.12	.95
g_2	.65	.84	.99	.94	.24	.99	.91	.14	.99	.95	.14	.95
g_3	.63	.85	.99	.93	.25	.99	.89	.15	.99	.94	.15	.95
Subsampling												
	Food			Clothing			Leisure			Transport		
Block	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3
b_1	.66	.88	.99	.99	.99	.99	.99	.94	1.0	.82	.13	.95
b_2	.55	.90	1.0	.99	.99	1.0	.92	.83	1.0	.54	.18	1.0

Certainly, the results from Table 3 need to be interpreted carefully, since the test is telling us that model is clearly separable. We do not know whether one of the regressors in the nonparametric part has a linear effect. Note that the results of testing additivity in Table 3 tell us that the model is additive separable in its nonparametric part, but they tell us nothing about the linearity of each component. In other words, it is possible to accept the nonparametric additive (separability) hypothesis even if one of those regressors has a linear effect on expenditure on different goods. The computed p-values concerned with testing linearity of each nonparametric component from model [13] are shown in Table 4. For this testing hypothesis procedure we use a bootstrap scheme for two bandwidth g_1=*1* and g_2=*1.2*. Again, we

define a set K_n (with cardinality L=*10*) of bandwidths k in a range from *0.35* to 2.

Note that for the clothing expenditure we are able to reject linearity of schooling at 10% for both bandwidths, and for the food expenditure we reject linearity of schooling at 7.9% (7.6% for g_2), in both cases with test statistic τ_1. For clothing expenditure, similar results on linearity of schooling are obtained with τ_2. With test τ_3 the percentage of rejection of linearity of schooling decreases to 6%. For the food expenditure, we are only able to reject linearity of age at 10% for both bandwidths. In the rest of the cases, we reject the linear effect hypothesis of age, schooling and expenditure at $\alpha \leq 5\%$ for all test statistics and for all bandwidths.

Table 4. Testing Individual linearity

		Age			Schooling			Expenditure		
		τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3
Clothing	g_1	.018	.014	.034	.10	.095	.062	.018	.05	0
	g_2	.016	.014	.028	.10	.10	.060	.016	.05	0
Food	g_1	.10	.002	.024	.079	.008	.020	0	0	0
	g_2	.10	.004	.020	.076	.008	.020	0	0	0
Leisure	g_1	.020	0	.004	.008	.030	0	0	0	0
	g_2	.010	0	.004	.008	.030	0	0	0	0
Transp	g_1	0	.018	0	.004	.002	.002	0	0	0
	g_2	0	.026	0	.004	.004	.002	0	0	0

Note that in general, linearity of age and schooling is rejected for every expenditure type. Moreover, for all test statistics and for all bandwidths the linear effect of total expenditure on expenditures categories is strongly rejected. From Tables 3-4 we conclude that the results are coherent with the shape of the curves estimated in Figures 1-4. This gives us an idea of the robustness and reliability of our methods.

4 Conclusions and Future Research

This paper applies semiparametric additive PLM regression techniques for studying the relationship between consumption and household characteristics based on the Spanish Expenditure Survey. On the one hand, in the case of clothing and leisure, the additive specification for nonparametric components

is (weakly) supported for test statistics based on errors of the additive PLM model and non,-semiparametric estimators, with the bootstrap scheme. However, with τ_1 and τ_3 test statistics we are unable to reject the null hypothesis of additivity for different resampling schemes. On the other hand, additive separable nonlinear effects are completely supported by the results on specification testing. In general terms, there is no evidence to assert that any linear effect of regressors of interest on the different expenditure categories is observed in the subsample SES data used in this analysis. In conclusion, the results from Tables 3-4 allow us to assert that the joint effect of total expenditure, age and schooling on expenditures categories is nonlinear additive separable.

The general results obtained from the estimation and testing of Engel curves show that modelling the effects of total expenditure on the different expenditure types simultaneously with other regressors such as those included here certainly deserves better treatment than usually found in one-dimensional semiparametric analysis. In particular we observe that households with younger heads tend to behave differently from other households, and clearly this fact is not captured in an Engel curve system in which only linear and quadratic age effects are included in the empirical specification.

Note that in this paper we only take into account a partial household composition (we only control for number of children, sex and number of adults). Therefore, a reasonable extension of empirical analysis with additive PLM (simple additivity does not allow such analysis) could be carried out by introducing more demographic variation to obtain variety in behavior (more regions, labor market, temporal dummy to capture price effects, etc.). Moreover, we could be interested in allowing Z_{ij} vary in any way with j and Stlusky symmetry, then would be necessary to impose a function to get general equivalence scale in order to fulfill conditions of proposition 5 in Blundell et al (2003).

Another interesting point to investigate is whether changes in consumer preferences take place over time and then to make an extension to dynamic models. One can take data from 1980 and 1990, for instance, and to make a comparison of consumer behavior. This would be an interesting question for the future. Finally, it would be interesting to extend the analysis to more categories of goods (health, furniture house, rent, etc).

References

[1] Arevalo, R., Cardelus, M. T., and Ruiz-Castillo, J. (1998) La Encuesta de Presupuestos Familiares de 1990-91. http://www.eco.uc3m.es/epf90-91.html.

[2] Banks, J. Blundell, R., and Lewbel, A. (1997). Quadratic Engel Curves and Consumer Demand. *The Review of Economics and Statistics*, **79**, No 4, 527-539.

[3] Barrientos-Marín, J and S. Sperlich (2006). The Size Problem of Kernel Based Bootstrap Test When the Null Is Nonparametric. Working in progress. University of Alicante.

[4] Bierens, H and H. Pott-Buter (1990) Specification of Household Engel Curves by Nonparametric Regression. *Econometric Reviews*, **9**, 123-184.

[5] Blundell, R., Duncan, A., and Pendakur, K. (1998). Seminparametric Estimation and Consumer Demand. *Journal of Applied Econometrics*, **13**, No 5, 435-461.

[6] Blundell, Richard; Browning, Martin and Ian A. Crawford (2003). Nonparametric Engel Curve and Revelead Preference. *Econometrica*, **71**, No 1, 205-240.

[7] Deaton, A and J. Muellbauer (1980a). An Almost Ideal Demand System. *American Economic Review*, **70**, 321-326.

[8] Deaton, A and J. Muellbauer (1980b). *Economic and Consumer Behavior*. Cambridge University Press, Cambridge.

[9] Dette Holger, C. Von Lieres and S. Sperlich (2004) A Comparison of Different Nonparametric Method for Inference on Additive Models. *Nonparametric Statistics*, **00**, 1-25.

[10] Gozalo, P. L. and O. B. Linton (2001) Testing Additivity in Generalized Nonparametric regression models with Estimated Parameters. *Journal of Econometrics*, **104**: 1-48.

[11] Härdle, W. and E. Mammen (1993) Comparing Nonparametric Versus Parametric Regression Fits. *Annals of Statistics*, **21**, No. 4, 1926-1947.

[12] Hardle, W., Huet, S., Mammen, E., and Sperlich, S (2004) Semiparametric Additive Indices for Binary Response and Generalized Additive Models. *Econometric Theory*, **20**, 265-300.

[13] Härdle, W., Müller, M., Sperlich, S., and Axel Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer-Verlag, 2004.

[14] Hengartner, N and S. Sperlich (2005) Rate Optimal Estimation with the Integration Method in the Presence of Many Covariates. *Journal of Multivariate Analysis*, **95**, Issue 2, 246-272

[15] Horowitz. J, L and V. Spokoiny (2002) An Adaptive, Rate-optimal Test of Parametric Mean-Regression Model Against A Nonparametric Alternative. *Econometrica*, **69**, No. 3, 599-631.

[16] Leser, C. E. V (1963). Form of Engel Functions, **31**, No 4, 694-703.

[17] Linton, O. B., and J. P. Nielsen. (1995). A Kernel Method of Estimating Structured Nonparametric regression Based on Marginal Integration. *Biometrika*, **82**, 93-101.

[18] Lyssiotou, P; Pashardes, P and Stengos, Thanasis (2001). Age Effects on Consumer Demand: An Additive Partially Linear Regression Model. *The Canadian Journal of Economics*, **35**, No 1, 153-165.

[19] Nadaraya, E. A. (1964). On Estimating Regression. *Theory Probability Applied,* **10**.

[20] Neumeyer, N and S. Sperlich (2005). Comparision of Separable Components in Different Samples. Workin Paper, Universidad Carlos III.

[21] Newey, W. K and J. Powell (2003) Instrumental Variables Estimation of Nonparametric Models. *Econometrica*, **71**, 1565-1578.

[22] Robinson, P (1988) Root N-Consistent Semiparametric Regression. *Econometrica*, **56**, 931-54.

[23] Rodriguez-Póo, J. M, S. Sperlich and P. Vieu (2005) And Adaptive Specification Test For Semiparametric Models. Working Paper.

[24] Sperlich, S. (2005). A Note on Nonparametric Estimation with Constructed Variables and Generated Regressors. Working Paper. Universidad Carlos III.

[25] Stone, C. J (1985). Additive Regression and Other Nonparametric Regression Models. *Annals of Statistics* **13**: 689–705.

[26] Stone, C. J (1986). The Dimensionality Reduction Principle for Generalized Additive Models. *Annals of Statistics* **14**, 592-506

[27] Watson, G. S (1964) Smooth Regression Analysis. *Sankhyā Ser.* A **26.**

[28] Working, H. (1943) Statistical Laws of Family Expenditure. *Journal of the American Statistical Association*, **38**, 4-56.