Food Expenditure Patterns of the Urban and the Rural Households in Greece. A Kernel Regression Analysis

Panos Fousekis and Panagiotis Lazaridis[1]

Abstract

Nonparametric (Kernel) regression analysis and micro-data from the Family Budget Survey (FBS) are used in this paper to estimate and to compare the Engel curves for food demand of the urban and the rural households. The empirical results suggest that the Characteristic Substitution Effects (CSEs) are not constant but vary considerably with the total consumption outlay. They also suggest that the Working-Leser hypothesis, according to which shares are linear in logarithmic expenditure, is consistent with the food demand patterns in Greece.

Keywords: *Engel Curves, Kernel Regression, Greece.*

Introduction

The share of total consumption expenditure allocated to food has been the focus of economic research since the 19th century. The observation that the demand for food increases at a rate lower than that of income (Engel Law) has led researchers to employ the food share as an (inverse) index of household welfare. The fact that this index does not depend on measurement units renders it appropriate for comparing the behavior of the same households over time or of different households in a given time period. In this sense, the factors determining allocations to food are important both for economic analysis as well as for policy formulation.

Until the late 1980s the relationship between food and total consumption expenditures was investigated via parametric techniques, exclusively. In other words, the researchers were making the assumption that the above mentioned relationship had a specific functional form which depended on a number of unknown parameters. The estimation of those parameters was leading subsequently to the determination of the Engel curve. Since then, in the relevant literature there has been a visible shift to nonparametric modeling (e.g. Atkinson et al., 1990; Banks, et al., 1997; Blundell, et al., 1993; Lewbel, 1991). The advantage of the nonparametric techniques is that they require no precise assumptions about the functional forms. Instead, they allow data to "speak for themselves", provi-

[1] Lecturer and Assistant Professor, respectively at the Agricultural University of Athens Address Correspondence to: Panagiotis Lazaridis, Department of Agricultural Economics and Development, Agricultural University of Athens, Iera Odos 75, Athens, Greece 118 55, Tel: 01-9524720, Fax: 01-9524786, E-mail: T.Lazaridis@aua.gr

ing in this way a robust approach to statistical inference which is more likely to capture the underlying structure. Moreover, the nonparametric models can serve as tools for assessing the validity of parametric specifications (Deaton, 1997; Pagan and Ullah, 1999; Yatchew, 1998). However, nonparametric techniques are theoretically complex, computationally intensive and require large data sets.

The objective of this paper is the estimation and comparison of the Engel curves for food demand in the urban and the rural areas of Greece. In contrast with earlier works on the topic (e.g. Veletzas and Karagiannis, 1993; Mergos and Donatos, 1989) the investigation here relies on nonparametric methods and, in particular, on Kernel regression analysis. In what follows section 2 presents the principles of Kernel regression and the Nadaraya-Watson estimator, while section 3 discusses the data. Section 4 includes the empirical application, comparisons among household types, and statistical tests on the Working-Leser hypothesis. Section 5 offers conclusions and suggestions for future research.

Nonparametric Regression and the Nadaraya-Watson Estimator

Econometric models describe the relationship between economic variables. This relationship is often represented by means of conditional moments. In particular, given Y a response variable and X an explanatory variable one is interested in estimating

$$m(x) = E(Y / X = x) \tag{1}$$

where $m(x)$ is the regression curve.

In parametric regression analysis $m(x)$ is assumed to follow a particular functional form (e.g. $m(x) = x'\beta$), where β is a vector of unknown parameters estimated using some loss function. The nonparametric regression analysis does not require any assumption about the form of $m(x)$ beyond that the regression curve is a smooth line. Relation (1) may be written as

$$E(Y / X = x) = m(x) = \int y \frac{f(x,y)}{f_x(x)}\, dy = \frac{\int y\, f(x,y)\, dy}{f_x(x)} \qquad (2),$$

where f_X is the marginal density function of X and f is the joint density function of X and Y. Given sample observations $\{X_i, Y_i\}$, i = 1, 2, …., n, the nonparametric estimator of f_X is

$$\hat{f}_x(x) = n^{-1} \sum_{j=1}^{n} K_h(x - X_j) \tag{3}$$

while the nonparametric estimator of the conditional expectation function $\int y f(x,y)\, dy$ is

$$\hat{r}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i \tag{4}$$

(Silverman, 1986; Yatchew, 1998; Haerdle, 1990; Haerdle et al., 1999).

In relations (3) and (4), $K_h = (1/h)K((x - X_i)/h)$ is a Kernel function with bandwidth, h.[1] Plugging in leads to the Nadaraya-Watson estimator at point $x \in X_1, \ldots, X_n$

$$\hat{m}_h(x) = \frac{\hat{r}}{\hat{f}_x} = \frac{1}{n}\sum_{i=1}^{n} \frac{h^{-1}K(\frac{x - X_i}{h})}{n^{-1}\sum_{j=1}^{n} h^{-1}K(\frac{x - X_j}{h})} Y_i = \frac{1}{n}\sum_{i=1} W_{hi}(x) Y_i \tag{5}$$

Given that $\frac{1}{n}\sum_{i=1} W_{hi}(x) = 1$, the Nadaraya-Watson estimator can be viewed as the weighted (local) average of the response variables Y_i around x. The Nadaraya-Watson estimator shares this weighted local average property with several other smoothing techniques (e.g. k-nearest neighbor, spline smoothing, and LOWESS-local regression method). The Nadaraya-Watson estimator, however, is the most commonly used in empirical economic analysis (Blundell, et al, 1993; Delgado and Miles, 1997; Banks et al. 1997; Yatchew, 1998). The parameter h determines the degree of smoothness of $\hat{m}(x)$. When $h \to 0$, then $m_h(X_i) \to Y_i$, meaning that we get an interpolation of the data. On the other hand, when $h \to \infty$, then $m_h(X_i) \to \bar{Y}$, meaning that the estimator is a constant function assigning the sample mean to each x. The Nadaraya-Watson estimator is consistent (Haerdle et al., 1999) and its variance can be calculated as

$$(\hat{\sigma}_h(x))^2 = n^{-1}[\sum_{i=1}^{n} W_{hi}(x)(Y_i - \hat{m}_h(x))^2] \tag{6}$$

The Data

The data for this study come from the 1993/94 Family Budget Survey (FBS) conducted by the National Statistical Service of Greece. The survey contains information on the quantities and on the money values of a very large number of goods and services consumed. It also contains information on certain socioeconomic and demographic characteristics. In 1993/94, a random sample of 6756 households were surveyed. For the purpose of the analysis the households in the sample were grouped using two criteria, namely, the area of residence and the number of children (less than 17 years old). On the basis of the first criterion, the distinction was drawn among couples with 0, 1, and 2 children. On the basis of the second criterion, it was drawn among urban (residing in cities with population at least 100000) and rural (residing towns and villages with population less than 10000). For all households selected, the age of the food planer lies between 20 and 65 years. In this way, 6 different types of households were created. Table

1 presents descriptive statistics on food shares, total consumption expenditure, and number of observations for the different types.

Table 1. Descriptive Statistics of the Data

Household Type	Number of Obs	Natural Logarithm of Total Expenditure		Budget Share (%)	
		(1)	(2)	(1)	(2)
Urban, no children	394	12.34	0.49	0.3	0.02
Urban, one child	282	12.64	0.33	0.27	0.01
Urban, two children	408	12.69	0.29	0.29	0.01
Rural, no children	377	11.98	0.55	0.37	0.03
Rural, one child	138	12.46	0.32	0.31	0.015
Rural, two children	403	12.61	0.34	0.33	0.02

(1): Sample Average, (2): Sample Variance. The expenditure figures are in 1993/94 Greek Drachmas.

The Empirical Results

a. Nonparametric Estimates of the Engle Curves

As in earlier empirical studies (e.g. Banks, et al., 1997; Blundell et al., 1993; Delgado and Miles, 1997) the Engel curves to be estimated are of the general form

$$S_i^l(X_i) = m^l(\ln(X_i)) + u_i \quad (7),$$

where i stands for the household, $l = 1, 2, \ldots, 6$ for its type, S for the food share, X for total consumption expenditure, and u for the disturbance term with zero mean and variance $Var(u_i / X_i) = \phi(X_i)$.[2]

The estimations have been performed in the econometric program *XploRe* employing the Quartic Kernel function

$$K_h = \frac{15}{16}(1 - v^2)^2 I(|v| \leq 1) \quad (8),$$

where I is an indicator variable and $v = (x - X_j)/h$, j=1, 2, …., n.[3] The bandwidth parameter h has been chosen optimally, that is, in such way as to minimize the asymptotical Mean Integrated Squared Error (MISE)[4]

$$MISE = \frac{1}{nh} + \frac{h^2}{12}\|f'_x\|_2^2 \quad (9),$$

where $\|f'_x\|_2^2$ is the L_2 – norm of the first derivative of the marginal density function of the independent variable.[5]

Figures 1 to 6 present the nonparametric estimates of the Engel curves along with the corresponding pointwise 95 percent confidence intervals. The precision of the estimates is very high for the central observations but it is lower in the tails. This behaviour is linked to so-called "boundary effect" which arises from the fact that at the boundary of the support only few observations can be averaged (Haerdle, 1990).[6] In all cases, the figures confirm the negative relationship between the food share and the total expenditure (Engel Law). Also, in most cases this relationship appear to be approximately linear (statistical tests on this are performed in 4c).

b. Characteristic Substitution Effects

Household characteristics are likely affect the share of a commodity in the total outlay (Delgado and Miles, 1997). Such change in the share (holding consumption expenditure constant) is called Characteristic Substitution Effect (CSE). Let a and b two household types. Then, the CSE is

$$CSE^{(a,b)}(x) = (\hat{m}^a(x)) - (\hat{m}^b(x)) \quad (10),$$

where symbol $\wedge$ stands for the nonparametric estimates of the respective regression functions. For the subsequent discussion a denotes urban while b denotes rural households.

Figures 7-9 present the CSEs for the urban and the rural households. In all cases the CSEs are negative for low levels of total consumption expenditure suggesting that, regardless of the number of children in a household, the poorer rural households tend to devote higher share of their budget to food compared to the poorer urban households. At higher levels of total consumption expenditure, however, the allocation patterns between food and non food items for the two types of households considered in this study are very similar. The result has reported in the study by Delgado and Miles (1997) for Spain. One may conclude, therefore, that as far as the allocation of total expenditure between food and non food commodities is concerned, the place of residence matters only for the poorer households.

The fact that the CSEs (especially for households with no and with two children) vary considerably with the level of total expenditure has important econometric implications.

Figure 1. Urban areas. Households with no children

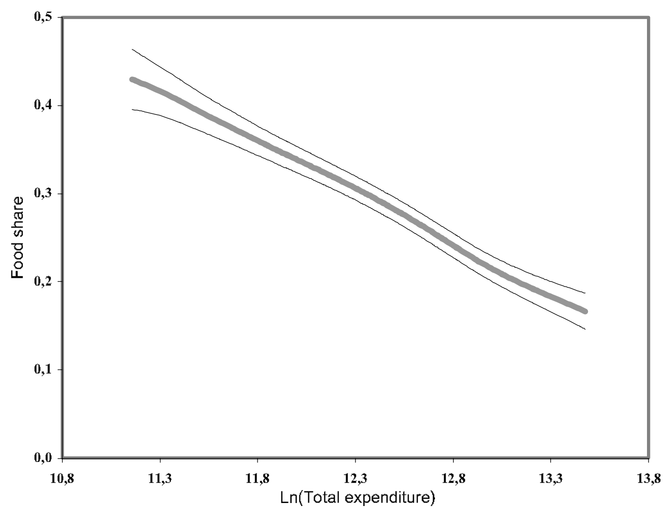


Figure 2. Urban areas. Households with one child

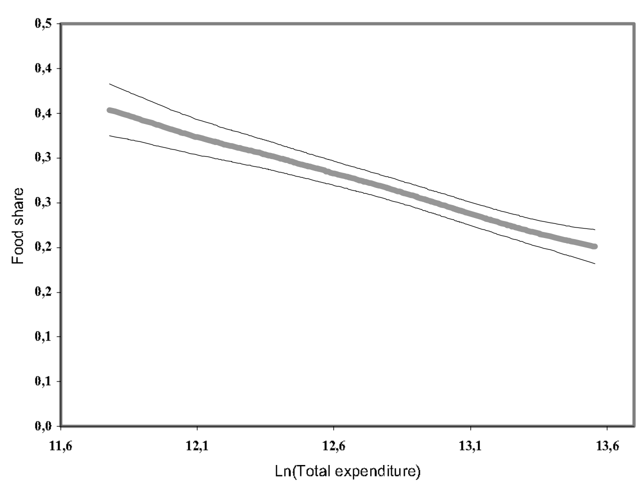


Figure 3. Urban areas. Households with two children

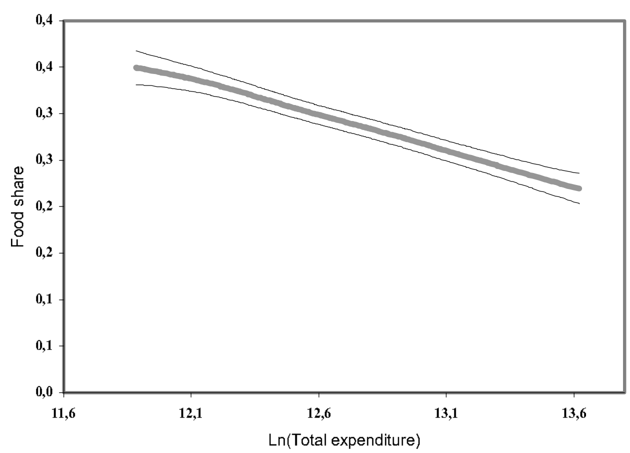


Figure 4. Rural areas. Households with no children

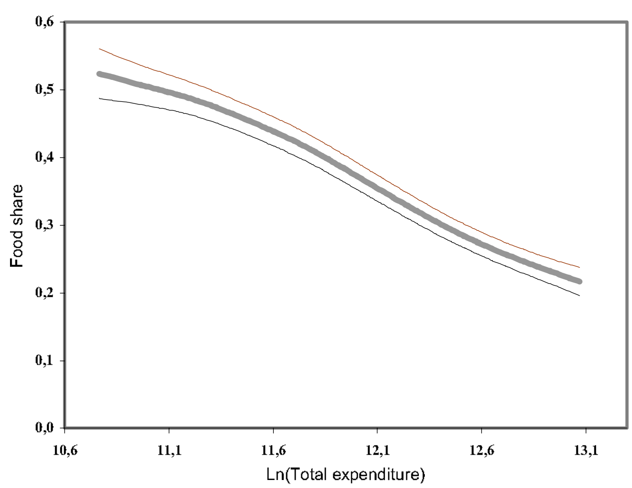


Figure 5. Rural areas. Households with one child

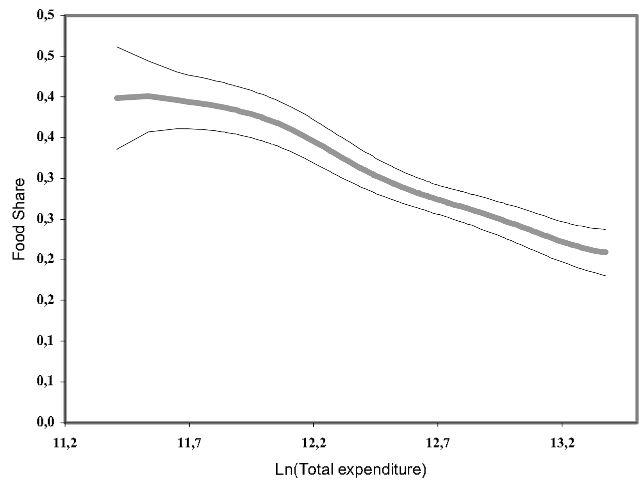


Figure 6. Rural areas. Households with two children

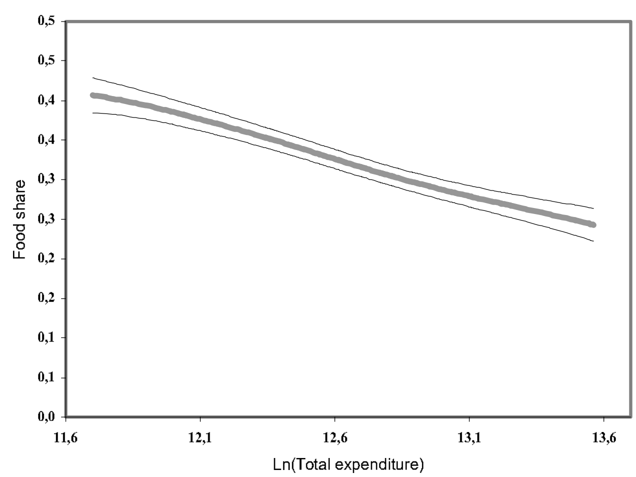


Figure 7. Characteristic Substitution Effect for Urban and Rural Households. Households with no children

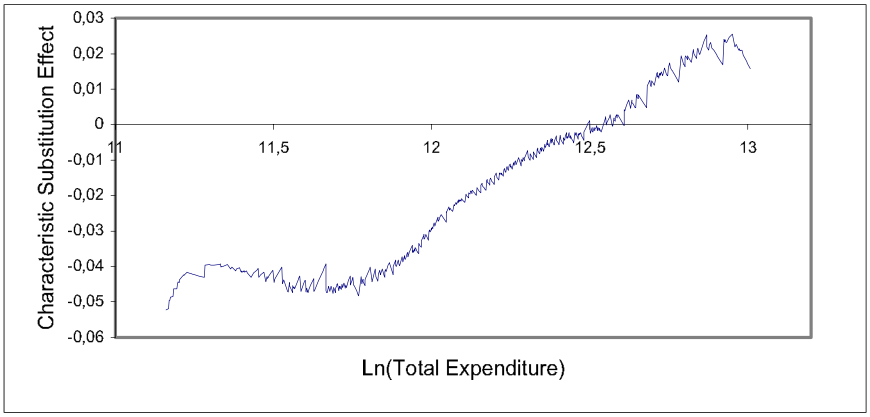


Figure 8. Characteristic Substitution Effect for Urban and Rural Households. Households with one child

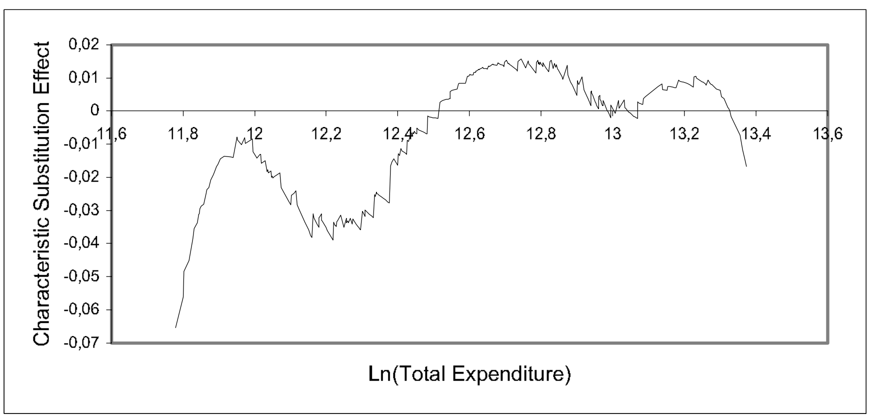
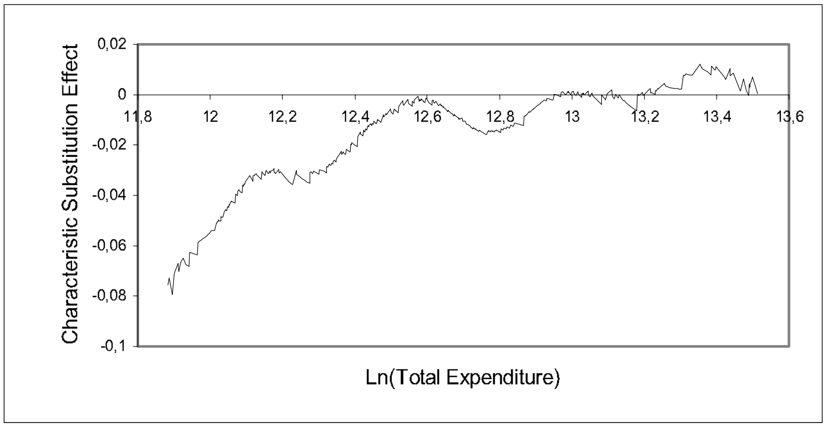


Figure 9. Characteristic Substitution Effect for Urban and Rural Households. Households with two children



These are that parametric models attempting to explain differences in food consumption behavior with the use of additive dummies are likely to be misspecified. The reason is that the additive dummies impose constant and independent of total consumption expenditure CSEs. The correct specification in this case requires the inclusion of the cross products of additive dummies and total consumption outlay.

c. A Nonparametric Test of the Working-Leser Hypothesis

A very common specification of Engel curves is the one in which the budget share of a commodity is a linear function of the natural logarithm of total expenditure (Working-Leser hypothesis). This Generalized Linear (GL) specification is a necessary and sufficient condition for aggregate demand functions to resemble representative agent models in certain ways (Muellbauer, 1975). It is also necessary for aggregate demands to exhibit the weak axiom of revealed preference (Freixas and Mas-Colell, 1987 refer to GL as the "no torsion" condition). The AIDS, the Translog, the CBS, and the Linear Expenditure models are GL ones and they are very frequently used in empirical analysis both because their tractability as well as because their exact aggregation or representative agent properties (Lewbel, 1991). All earlier empirical works on food demand in Greece relied on the Working-Leser hypothesis.

In the recent years, however, the empirical research found evidence that this hypothesis is not always valid. With regard to the shape of Engle curves for food the existing evidence is mixed. Lewbel (1991), Banks, et al. (1997), and Delgado and Miles (1997) conclude that the Working-Leser hypothesis holds for the UK and Spain. Moro and Sckokai (2000), however, report that quadratic terms in logarithmic income are required in Italy.

In parametric settings specification tests are usually performed considering nested or non-nested models. These tests, however, are only consistent in the direction of certain parametric alternatives. When the Working-Leser hypothesis is the issue one wishes to test

$$H_0: \quad E(S_i^l) = a^l + b^l \log(X_i)$$
$$H_1: \quad E(S_i^l) = H^l(\log(X_i))$$

where a^l and b^l are the parameters of the parametric GL model and H is an unknown smooth function. The validity of the Working-Leser hypothesis for food demand in the rural and the urban areas of Greece is assessed in this paper using the HH-test which is consistent in the direction of general nonparametric alternatives (Horowitz and Haerdle, 1994). The HH-test statistic is

$$T = h^{0.5} \frac{\sum_{i=1} (S_i - \hat{a} - \hat{b}\log(X_i))\,(\hat{m}(\hat{a} + \hat{b}\log(X_i)) - (\hat{a} + \hat{b}\log(X_i)))}{n^{-1} \sum_{i=1} \{S_i - \hat{m}(\hat{a} + \hat{b}\log(X_i))\}^2} \quad (11),$$

where h is a bandwidth parameter, $\hat{m}$ is the nonparametric regression function of the budget share on the predicted values from the parametric model, and n is the number of observations. The null hypothesis (that is, the Working-Lesser specification) can be accepted or rejected at level ζ according to whether the value of the test statistic exceeds the $1-\zeta$ quantile of the standard normal distribution, being an one sided test.[7] Table 2 presents the results of the application of the HH-test to the 6 household types. In all cases the null hypothesis cannot be rejected. This suggests that the Working-Lesser hypothesis provides an adequate specification of the food Engel curves in Greece.

Table 2. HH-tests of the Working-Leser Hypothesis

Household Type	**Empirical Value of the HH Statistic***
Urban, no children	0.47
Urban, one child	0.09
Urban, two children	0.45
Rural, no children	0.88
Rural, one child	0.25
Rural, two children	0.55

* Theoretical value at the 5 percent level is 1.65 (one-sided test)

d. Average Derivative Estimates and Parametric Slopes

In parametric models, we are concerned with the estimation of slopes which give the response of the expected value of the dependent variable to a marginal change in the independent variable. In nonparametric models, the same information can be achieved through the Average Derivative Estimator (ADE). The ADE is defined as

$$\delta_h(x) = E(\frac{\partial m_h}{\partial x}) \qquad (12)$$

(Haerdle και Stoker 1989; Stoker 1991) where m is the regression function and x is the independent variable. As it is the case with the regression function itself, the ADE depends also on the bandwidth parameter h. Haerdle and Stocker (1989) show that the Average Derivative can be computed as

$$\hat{\delta}_h(x) = n^{-1}\sum_{i=1} \hat{\phi}_h(x)Y_i \qquad (13).$$

In (13), $\hat{\phi}_h$ is the estimate of the "score function", that is, of the ratio $-\frac{f_x'}{f_x}$ where f_x is the marginal density of the independent variable, and $(')$ denotes its first derivative.

Table 3 reports a comparison between the Average Derivatives $(\hat{\delta}^l)$ of the nonparametric model and the slopes $(\hat{b}^l)$, l =,1, .., 6, estimated parametrically under the null (Working-Leser specification). The slopes of the parametric models are very similar to the ADEs. The standard errors of the later, however, are substantially higher than those of the corresponding parametric slopes. There are two explanations for the differences. First, when a parameteric model is a correct specification of the underlying relationship then the slope estimates converge more rapidly to the true parameters than the ADEs. This is the case here since the HH-test (section 4c) failed to reject the parametric null hypothesis.

Table 3. Average Derivatives and Parametric Slopes*

Household Type	Parametric Slope	Average Derivative
Urban, no children	-0.148* (0.009)	-0.142* (0.035)
Urban, one child	-0.115* (0.012)	-0.098* (0.041)
Urban, two children	-0.1* (0.009)	-0.099* (0.04)
Rural, no children	-0.182* (0.011)	-0.163* (0.07)
Rural, one child	-0.136* (0.01)	-0.137* (0.06)
Rural, two children	-0.133* (0.011)	-0.128* (0.051)

* Statistically significant at the 5 percent level or less. Standard errors in parentheses. The standard errors for the Average Derivatives have been calculated as in Haerdle and Stoker (1989, pp. 998-99)

The higher rate of convergence implies higher efficiency of the slope estimates something which is translated into smaller standard errors; Second, the non parametric techniques do not require that marginal effects to be constant over the domain of the independent variables allowing, thus, for greater variability which, in turn, is reflected in the standard errors (Delgado and Miles, 1997; Pagan and Ullah, 1999; Yatchew, 1998).

Both the slopes and the ADEs change substantially with household types. This confirms the conclusion reached from the examination of the CSE (in section 4b) that demographic and socioeconomic factors do not cause parallel shifts in the Engel curves of food demand and, thus, the use of additive dummies is not likely to ensure an adequate specification for a model relying on the total sample (all household types).

The expenditure elasticities of demand for food for each household type may be calculated as $(1+\frac{\psi^l}{S^l})$ where ψ^l stands for the estimate of the ADE (slope) from the nonparametric (parametric) model. On the basis of the ADEs, the expenditure elasticities range from 0.52 (for urban households with no children) to 0.65 (for urban households with two children). On the basis of the parametric slopes they range from 0.5 (for urban households with no children) to and 0.65 (for urban households with two children). These results are very similar to those reported by Veletzas and Karagiannis (1993) and Mergos and Donatos (1989).

Conclusions

The objective of the present study was the estimation and the comparison of the Engel curves for food demand in the urban and the rural areas of Greece. In contrast with past works on the topic, the investigation here relies on non-parametric methods and, in particular, on Kernel regression analysis. The main results may be summarized as follows:

a) The food share (holding the total consumption expenditure and the number of children constant) is higher for rural households. There is a clear trend for convergence, however, as the total outlay increases. Given that total expenditure is a proxy for household disposable income one may conclude that differences in behavior with respect to food demand are likely to be found between the poorer urban and rural households rather than between the wealthier ones. Those differences could be attributed to the relatively higher amount of consumption from own production of basic food items in rural areas, to relatively higher availability of other consumer goods (education, health, entertainment etc) in urban areas and to higher rents (real or imputed) paid in urban areas.

b) The Working-Leser hypothesis, according to which the budget share is a linear function of the logarithmic income, cannot be rejected for all household types. This finding is consistent with the results of Delgado and Miles (1997) for Spain, of Banks et al. (1997) and Lewbel (1991) for the UK. It contrasts, however, with the findings of Moro and Sckokai (2000) for Italy. The results of the present study suggest the parametric models used in past studies to analyze demand for food in Greece were correctly specified.

c) The Characteristic Substitution Effects are not constant but vary considerably with total consumption outlay suggesting that additive dummies do not ensure an adequate specification in models attempting to explain differences in food consumption behavior among household types.

The present paper has focused on food demand. Food, however, is a composite commodity and aggregation may sometimes obscure the characteristics of the Engel curves. Therefore, an empirical investigation at a lower level of aggregation (e.g. for simpler food commodities such as meat, dairy, vegetables etc) is certainly warranted.

Notes

1. Kernels are probability density functions which are symmetric around zero.
2. It is not necessary that the variance of u is a constant function. Typically one assumes that ϕ is a continuous and bounded function (Haerdle et al., 2000).
3. There are many possible choices for the Kernel function such as Epanechnikov, Triangular, Gaussian etc. The relative literature suggests that choice of the Kernel is not a critical one (e.g. Deaton, 1997; Haerdle et al, 1999; Haerdle et al, 2000). The theoretical background for this suggestion is that Kernel functions can be rescaled such as the difference between two kernel density estimates using two different Kernels is almost negligible (Marron and Nolan, 1988). In applied work, Kernels are chosen on the grounds of computational convenience and/or differentiability at the boundary. The quartic (biweight) Kernel employed here offers computational convenience and its derivative is continuous at the end of the band.
4. The MISE is the integral over the squared distance between the true and the estimated density function.
5. For further details on optimal bandwidth selection see Silverman (1986), Haerdle (1990) and Wand and Jones (1995).
6. We thank an anonymous reviewer for pointing it out. For kernels with automatic adaptability for boundary effects near endpoints see Zhang and Karunamuni (1998).
7. The test is one sided since T diverges to $+\infty$ under the alternative hypothesis against which it is consistent.

References

Atkinson, A., Gomulka, J. and Stern, N. (1990). Spending on Alcohol: Evidence from the Family Expenditure Survey, 1970-83. Economic Journal, 100, 808-27.

Banks, J., Blundell, R. and Lewbel, A. (1997). Quadratic Engel Curves and Consumer Demand. The Review of Economics and Statistics, LXXIX, 527-39.

Blundell, R., Pashardes, P. and Weber, G. (1993). "What do we Learn about Consumer Demand Patterns from Micro-data." American Economic Review, 83,570-597.

Deaton, A. (1997). The analysis of Household Surveys: A Micro-econometric Approach to Development Policy. John Hopkins, Baltimore.

Delgado, M. and Miles, D. (1997). Household Characteristics and Consumption Behavior. Empirical Economics, 22, 409-29.

Freixas, X. and Mas-Collel, A. (1987). Engel Curves Leading to the Weak Axiom in the Aggregate. Econometrica, 55, 515-32.

Haerdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Haerdle, W. and Stoker, T. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. Journal of the American Statistical Association, 84, 986-95.

Haerdle, W., Klinke, S.and Muller, M. (2000). Xplore Learning Guide. Springer.

Haerlde, W., Muller, M., Sperlich, S.and Werwatz, A. (1999). Non and Semiparametric Modeling. Humboldt-Universitat zu Berlin, Berlin.

Horowitz, J and Haerdle, A. (1994). Testing a Parametric Model Against a Semiparametric Alternative. Econometric Theory , 10, 821-48.

Lewbel, A. (1991). The Rank of Demand Systems: Theory and Nonparametric Estimation. Econometrica, 59, 711-30.

Marron, J. and Nollan, D. (1988). Canonical Kernels for Density Estimation. Statistics and Probability Letters, 7, 195-99.

Mergos, G., and Donatos G. (1989). Demand for Food in Greece. An Almost Ideal Demand System Analysis. Journal of Agricultural Economics, 40,178-84.

Muellbauer, J. (1975). Community Preferences and Representative Consumer. Econometrica, 44, 525-43.

Pagan, A., and Ullah A. (1999). Nonparametric Econometrics, Cambridge University Press. Silverman, B. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall.

Stoker, T. (1991). Equivalence of Direct, Indirect, and Slope Estimators of Average Derivatives, in: Barnett, Powel, and Tauchen (eds), Nonparametric and Semiparametric Methods inEconometrics and Statistics. Cambridge University Press.

Yatchew, A. (1998). Nonparametric Regression Techniques in Economics. Journal of EconomicLiterature, XXXVI, 669-721.

Veletzas, K. and Karagiannis, G.(1993). An Empirical Analysis of Demand for Consumer Goods in Greece: 1958-89. Epistimoniki Epetirida of Macedonia University, 11, 94-114 (in Greek)

Wand, M. and Jones, M. (1995). Kernel Smoothing. Chapman & Hall.

Zhang, S., and Karunamuni, R. (1998). On Kernel Density Estimation Near Endpoints. Journal of Statistical Planning and Inference, 70, 301-316.