Specification of household expenditure functions and equivalence scales by nonparametric regression

Bierens, H.J.; Pott-Buter, A.

1987

SERIE RESEARCH MEMORANDA

SPECIFICATION OF HOUSEHOLD EXPENDITURE FUNCTIONS AND EQUIVALENCE SCALES BY NONPARAMETRIC REGRESSION

Herman J. Bierens
Hettie A. Pott-Buter

Research Memorandum 1987-44 Dec. '87

VRIJE UNIVERSITEIT
FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
EN ECONOMETRIE
AMSTERDAM

SPECIFICATION OF HOUSEHOLD EXPENDITURE FUNCTIONS AND EQUIVALENCE SCALES BY NONPARAMETRIC REGRESSION *)

by

Herman J. BIERENS

Free University, Department of Econometrics, 1081 HV Amsterdam

and

Hettie A. POTT-BUTER

University of Amsterdam, Department of Economics, 1011 NH Amsterdam

Abstract: This paper demonstrates the usefulness of nonparametric regression analysis for functional specification of household expenditure functions and equivalence scales without restricting the class of admissible functional forms, using a Dutch budget survey.

1. INTRODUCTION

Household expenditure patterns differ across households according to family size, age composition, educational levels and other household characteristics. In modelling household expenditures one should therefore not only relate expenditures to income and commodity prices, but also to these household characteristics. These models then form a basis for welfare comparison between households and the estimation of the cost of children, directly or indirectly via the construction of household equivalence scales.

Research in this area is of considerable practical significance. Knowledge of the cost of children is necessary, for example, for judges who have to assess alimony and politicians who decide on the level of child and family allowances. Most industrial countries have a system of family allowances compensating the direct cost of children. These allowances however are seldom based on the result of advanced academic studies. In particular the econometric studies appear to have had little or no impact on social policy. Only the results of the more traditional methods which are also relatively easy to understand seem to have had any influence. These simple methods, however, lack in general theoretical foundation and an objective

---

*) The comments of Wim Groot and Arie Kapteyn are gratefully acknowledged.

base.

The various econometric approaches to estimating household expenditure functions and household equivalence scales all have one other major drawback in common, namely that the functional form of the demand equations has to be specified in advance, directly or indirectly via the specification of the functional form of the utility function. The functional form of the model or the utility function is usually chosen on the basis of tractability rather than on the basis of a priori knowledge of the true functional form. Tractability and reality, however, need not coincide in practice. Since there is almost a continuum of theoretically admissible functional forms, the actually chosen functional form is almost surely misspecified. This situation is reminiscent of drawing a random variable from a continuous distribution, i.e., the probability that this random variable equals a certain fixed value is equal to zero.

Misspecification of the functional form of household expenditure functions may have serious consequences for the econometric results. In particular, functional misspecification usually leads to inconsistent parameter estimators, and consequently the estimated equivalence scales are inconsistent too. In this paper we are mainly concerned with the functional specification of household expenditure systems. After a review of the literature on demand functions and equivalence scales in Section 2 and the functional specifications used, we derive in Section 3 the functional form of our expenditure functions from nonparametric regression results, using the 1980 Budget Survey for the Netherlands, in order to avoid model misspecification. Nonparametric regression analysis is a technique which allows consistent estimation of a regression model without specifying in advance its functional form. Thus the model is derived directly from the data, without restricting its functional form. The only specification that is involved concerns the choice of the dependent variable and the independent variables. The relationship between this dependent variable and the independent variables is left free, apart from some mild regularity conditions (such as continuity). The nonparametric regression results are then translated to suitable parametric functional specifications, i.e., we have chosen parametric functional forms in accordance with the nonparametric regression results. These parametric specifications have been estimated by least squares, and various parameter restrictions have been tested in order to

simplify the models.

2. REVIEW OF THE LITERATURE

The methods of estimating household equivalence scales and the costs of children can be divided in:

a) Income evaluation methods.
b) Normative or 'minimum basket of commodities' methods.
c) Budgetary and econometric methods, on the basis of household expenditure surveys.

The income evaluation method

The income evaluation approach for assessing the cost of children is based on the answer to the so-called income evaluation question:
"Please try to indicate what you consider to be an appropriate amount of money for each of the following cases? Under my (our) conditions I would call an after-tax income per week (month/year) of about ..... very bad, of about ..... bad, of about ..... insufficient, of about ..... sufficient, of about ..... good, and of about ..... very good."
The response is related to the age and socio-economic category of the head of the household, his/her past and present income and the household composition (Kapteyn and Van Praag 1976). However, the method has been criticised on its assumptions, in particular the existence, measurability and interpersonal comparability of the individual welfare functions.

The 'minimum basket' method

A traditional method to estimate the direct minimum cost is to draw up detailed lists with minimum requirements of food, clothing, housing and so on for different types of households. The 'minimum basket' was originally based on a biological subsistence minimum, but it is now recognised that these estimates cannot be absolute and that only normative and relative packets for social welfare levels of average households can be established. Internationally the studies of Rowntree (1901, 1941) for establishing minimum subsistence levels for households of different composition have

received much attention. Rowntree's first 'minimum basket' was based on minimum diet estimates by Atwater (1895) and Atwater and Wood (1896), diets without meat and scarcely enough to live on. Yet these estimates were higher than most other nutritional scales.*) Rowntree's findings were used by Beveridge (1942). Beveridge's recommendations form the basis of the since 1948 operative social benefit system in England. Variants of Rowntree's approach are those of Oishansky (1965, 1968). He multiplied the food packet by an estimated value of the average income-food ratio.

Econometric methods

The budgetary (or econometric) approaches, i.e. empirical investigations of the expenditure behavior of households, are usually based on consumer theory. The theoretical basis has been improved with the empirical progress (see also Muellbauer 1977) but major drawbacks still exist. The different approaches and different models embody different conceptions of child cost and this can lead to quite different measures of the cost of children or equivalence scales (see also Deaton and Muellbauer 1986). Costs are by definition equal to the value of what is sacrificed. For the daily care and upbringing of children not only money is 'sacrificed' for food, clothing etcetera (direct costs) or education (party collective costs) but also time (cost of care) and immaterial costs (sorrow, worries). Most economic literature focus on the direct cost only. The other costs and the benefits of children are usually neglected. Traditionally, household composition is treated as exogeneous and the utility level of the household defined on current consumption. The utility derived from children is untill recently (e.g., Blundell and Walker 1982, 1984) ignored. The same applies to the forgone (mainly female) time available for work or leisure due to the presence of children. Collective costs are neglected in all the econometric studies (to the best of our knowledge) and only total

---

*) The Atwater scale belongs to the male-equivalent scale. A 17 year male = 100 (female = 80), male 14, 15, 16 = 80 (female 70), child 10, 11, 12, 13 = 60, child 6, 7, 8, 9 = 50, child 2, 3, 4, 5 = 40 and babies of 0, 1 = 30. This scale is more or the less identical to the 'König'scale of 1882, but higher than other well-known German scales (Nasse 1891 and Kuhna 1894), the Denmark scale (1897), the Swedish scale (1908), and the often applied Amsterdam scale (1917).

expenditure out of net household income is considered. Moreover, in general specific expenditures are related to total expenditure rather than household income, in order to impose the budget constraint. In addition most studies neglect life cycle influences and relative income and power distribution within the household (an exception is Bojer 1977).

The founding-father of household expenditure analysis is Engel (1883, 1895). Engel's equivalence scales are based on the proportion of income used for food of Belgium factory workers. The method assumes that the welfare of two households is equal if they spend the same proportion of their income on food. The equivalent scale m_0 depends on household composition. If $p_i q_i$ is the household expenditure on good i and x is household income then $p_i q_i / m_0$ is a function of x/m_0. Following Engel the first applicants of the method used equivalence scales m_0 based on nutritional requirements determined by experts [Stone (1954) used the Amsterdam scale of 1917] and total expenditure instead of household income. Although the theory is rather restrictive, as the equivalence scale is the same for each commodity, the method has been widely used since the beginning of this century all over the world.

Muellbauer (1977) estimated scales with Engel's method using British data from the Family Expenditure Survey under the hypothesis that the equivalence scales take the form

$$m = 1 + \delta_1 a_1 + \delta_2 a_2 ,$$

where δ_1 and δ_2 are parameters, a_1 is the number of children in the age group 0-4 and a_2 the number of children in the age group 5-16.

A specification of the Engel function that frequently fits the data well is the Working (1943) - Leser (1963) form, in which the food share w_f is a linear function of the logarithm of total expenditure. A simple extension that incorporates demographic effects is chosen by Deaton and Muellbauer (1986):

$$w_f = \alpha - \beta \ln(x/n) + \sum_{j=1}^{J} \gamma_j n_j + \varepsilon$$

where n_j is the number of persons in category j (j=1,..,J), n is the total number of persons in the household, x is total expenditure, α, β and the

γ_j's are parameters and ε is a random error. For many third world surveys Deaton and Muellbauer (1986) found that the $\ln(x/n)$ term provides a high degree of the explained variation and that the γ parameters are rather small.

Sydenstricker and King (1921) were the first to envisage the possibility of incorporating household composition as a variable in Engel curves by weighting the specific equivalence scales for particular commodities. A similar approach, independently discovered from Sydenstricker and King is followed by Prais (1953) and Prais-Houthakker (1955). The Prais (1953) and Prais and Houthakker (1955) model generalises the Engel model by allowing different demographic effects for each commodity and assumes Marshallian demand functions of the form

$$q_i/m_i = f_i(p, x/m_0), \quad i=1,2,..,k,$$

where q_i is the demand of commodity i, p is a k-vector of prices, x is total expenditure, m_i is the commodity-specific equivalence scale of commodity i and m_0 is the general (income) equivalence scale. The commodity-specific equivalence scales are functions of household composition only. The general or income coefficients can be expressed as a function of the specific commodity scales, because the exhaustive set of Engel curves must satisfy the budget restriction. Since the income scale can be expressed in terms of the specific scales it appears that only the latter need to be estimated. However, it is impossible to estimate the complete set of specific equivalence scales. Prais and Houthakker proposed an iterative procedure, but they did not put it into practice. It was left to Forsyth (1960), who set out to complete the work, to discover that the specific scales cannot be identified. Cramer (1969) summarises the main argument by means of a simplified example and concludes "The only way to remedy this situation is to impose yet another restriction on some or all of the coefficients in order to ensure that the set of equivalent adult scales is determinate. Thus Prais and Houthakker succeed in estimating the specific coefficients because they assume at the very outset that the income coefficients are unity for all individuals" (Cramer, 1969, p.168). The authors, using semi- and double logarithm Engel curves applied to pre-war British data, did not seem aware that they would be unable to estimate the specific

coefficients without the implied imposed restriction.

A solution to the problem of indeterminacy is a technique originally suggested by Rothbarth (1943) but in general named after one of the first applicants: Nicholson (1949). He estimates the income coefficients of children by considering commodities for which the child's specific coefficients may be reasonably fixed at zero. Other applicants include Henderson (1949), Dublin and Lotha (1974) and Deaton (1981).

The Engel method calculates the amount of money that would restore the previous food share, the Nicholson method the amount to restore the previous level of expenditures on adult goods. So the Nicholson method assumes that households of different sizes enjoy the same standard of living as long as the expenditures on a so-called representative basket of goods, which parents only acquire for themselves, per type of household is the same. Nicholson selected men's and women's clothing, tobacco and alcoholic beverages as commodities for which the specific equivalence scale values for children could be expected to be zero. Cramer (1969) argues that this approach provides the only justifiable solution to the problem of indeterminacy, but that unfortunately the empirical results are disappointing, largely because the commodities (e.g. alcoholic beverages, tobacco) concerned are liable to larger disturbances as well as observational errors than others. Cramer (1969, p.169) borrows the results of Forsyth (1960) for double logarithmic Engel curves of equal slope but varying intercept, to consider the effects of 1, 2 and 3 children on expenditure on alcoholic beverages, tobacco and entertainment, with rather disappointing results.

The Engel and Nicholson methods make different and mutually incompatible assumptions about the nature of the cost of children. Deaton and Muellbauer (1986) argue that under mild assumptions the Engel method produces estimates that are too large and the Nicholson (Rothbarth) method, though more plausible, estimates that are too small. Under more restrictive assumptions they derive a system of inequalities linking the two measures with more general measures based on Gorman's (1976) extension of the model of Barten (1964). The Barten/Gorman costs are in between the Engel and Nicholson estimates (Deaton and Muellbauer, 1986). The Prais-Houthakker model has not only been criticized by Forsyth (1960) and Cramer (1969) regarding the identification of the equivalence scales, but also by Muellbauer (1980) who argues that if the model is interpreted in terms of

utility theory it is consistent with a Leontief utility function only, hence no substitution between commodities is possible. The Barten/ Gorman model can be regarded as a generalisation of the previous models.

In Barten's (1964) model the Marshallian demand functions take the form

$$q_i/m_i = f_i(x/(p_1m_1),..,x/(p_km_k)),$$

where x is total expenditure, p_i is the price of commodity i and m_i is the corresponding specific equivalence scale. From the form of the Marshallian demand function it is obvious that a change in household composition has two effects, a direct effect through m_i and an indirect effect through the terms x/p_im_i (a pseudo price change substitution effect). Barten examined the case where the functions have a form which may be regarded as a variant of the 'Rotterdam School demand models'. This model is consistent with utility theory. However, estimation of Barten's model requires price information and hence pooled data, in order to prevent identification problems. This model has been applied by Blundell (1980), Brown and Deaton (1972), Bojer (1977), Deaton and Muellbauer (1980), Gorman (1976), Muellbauer (1975, 1977), Pollak and Wales (1981) and Ray (1985).

A disadvantage of Barten's model is that it assumes an excessive substitution effect as a result of changes in the household composition. Morover, there are important types of behavior that the model cannot accommodate. In particular, if the reference household (without children) does not consume the good, neither will the household with children except through the operation of substitution effects. This is not consistent with the Barton formulation except under extremely farfetched assumptions about substitution. See Deaton and Muellbauer (1986). Gorman's (1976) modification solves this problem by adding fixed cost of children to the Barten cost function.

Ray's (1983) general equivalence scale m_0 relates the cost function c^H of household H with z children and utility level u to the cost function c^R of a reference household with no children:

$$c^H(u,p,z) = m_0(z,p,u)c^R(u,p),$$

where p is the price vector. Choosing as a functional specification for m_0 and c^R:

$$m_0(z,p,u) = e^{\varepsilon_1 z+\varepsilon_2 z^2} \Pi_k p_k^{\delta_k} z e^{\lambda uz}$$

and

$$\log c^R = \alpha_0 + \sum_{i=1}^{n}\alpha_i \log p_i + \tfrac{1}{2}\sum_i\sum_j \gamma_{ij}(\log p_i)(\log p_j) + u\beta_0\Pi_k p_k^{\beta_k},$$

respectively, yields the (AIDS) cost function and the corresponding Marshallian demand system. In Ray (1985) a particular version of Barten's model is nested in this framework and tested against Ray's approach, resulting in a rejection of this particular Barten model. Ray chooses two general forms of the indirect utility functions, both giving rise to demand systems which allow non-separable preferences and non-linear Engel curves. The first is the Non-Linear Preference System, which lets (partly) the data determine the extent of non-linearity of the Engel curve. The second is the Almost Ideal Demand System (AIDS), poposed by Deaton and Muellbauer (1980), with children included as proposed in Ray (1983).

A new approach to the problem of demographic specification is suggested by Blundell and Walker (1984). Their approach borrows the household production framework from the neoclassical fertility literature where children might yield utility, but as in the traditional demand literature, children are assumed to be predetermined or rationed in the observed data. Thus a given demographic structure requires the input of market goods and time to maintain it at its given level. The household's problem is described as minimising the full expenditure required to attain a given level of utility $\bar{u}$ subject to the full income constraint and the household production function. The household's full expenditure function is defined by

$$C(p,z,\bar{u}) = C_1(p,z,\bar{u}) + C_2(p,z), \qquad (2.1)$$

where C_1 is the cost function associated to household demand of consumption goods and C_2 the cost function associated to household demand of goods used in household production to maintain the given demographic structure z, with p a vector of given prices (and wages) and z the number of children. Demo-

graphic variables have two effects, an income effect via $C_2(p,z)$ and a substitution effect via $C_1(p,z,\bar{u})$. Blundell and Walker use the nonseparable Gorman Polar Form as a specification of the households's consumption expenditure function:

$$C_1(p,z,\bar{u}) = A(p) + B(p,z)\bar{u} \tag{2.2}$$

where

$$A(p) = \sum_i \sum_j \alpha_{ij} p_i^{\frac{1}{2}} q_j^{\frac{1}{2}}, \quad \text{and} \quad B(p,z) = \Pi_i p_i^{\beta_i(z)}, \quad i,j=1,\ldots,n.$$

The specification employed for the cost function $C_2(p,z)$ corresponding to the household production function is the Generalised Leontief due to Diewert (1971):

$$C_2(p,z) = \sum_i \sum_j \gamma_{ij}(z) p_i^{\frac{1}{2}} p_j^{\frac{1}{2}}, \tag{2.3}$$

where $\gamma_{ij}(z)$ is a linear function of z under constant returns and $\gamma_{ij}(z)=0$ for $i \neq j$ corresponds to fixed coefficients. Adding (2.2) to (2.3) yields the full expenditure function (2.1), whose expenditure share equations can be written as

$$w_i = \sum_j [\alpha_{ij}+\gamma_{ij}(z)] r_i^{\frac{1}{2}} r_j^{\frac{1}{2}} + \beta_i(z)\{1-\sum_i \sum_j [\alpha_{ij}+\gamma_{ij}(z)] r_i^{\frac{1}{2}} r_j^{\frac{1}{2}}\} \tag{2.4}$$

where $w_i = r_i q_i$, $r_i = p_i/y$ and y is (full) expenditure. Adding up requires that $\sum_i \beta_i(z) = 1$ and symmetry requires that $\alpha_{ij} = \alpha_{ji}$, all $i=j$. Since (2.4) is a nonseparable generalisation of the LES they refer to it as NLES. Data from UK Family Expenditure Surveys for 1968 to 1981 were used for estimation. Three categories of expenditures are analysed: food, clothing and energy.

Main purpose of Blundell and Walker's (1982) earlier article is testing the commonly assumed restriction on the household's preferences of (weak) separability between goods and leisure. This restriction has been rejected. They introduce demographic variables to capture the effect of household composition not only on commodity demands but also on labour supply. The method used is an extension of the translation approach of Pollak and Wales (1978) to the leisure goods model, explored in more detail

in Blundell (1980).

3. SPECIFYING HOUSEHOLD EXPENDITURE FUNCTIONS BY NONPARAMETRIC REGRESSION

3.1. *Introduction*

A serious problem in all econometric approaches is the arbitrariness of the functional specification of the equivalence scales and the demand or cost functions. Although utility theory imposes certain restrictions on the functional specification of demand and cost functions, the class of theoretically admissible functional forms is almost uncountably large. Usually the functional form is chosen as to facilitate estimation rather than to approximate reality, so that all the models considered in the literature are likely misspecified.

Actually, all approaches ultimately amount to direct or indirect functional specification of the Marshallian demand functions as known functions of prices p, income or total expenditure x, household composition z and unknown parameters. Thus, denoting $q_i = g_i(p,x,z)$, i=1,..,k, the Marshallian demand functions, the various methods distinguish themselves by different recipes for the specification of the functions g_i.

In this paper we follow a different approach by estimating these functions g_i directly from the data, without specifying in advance any functional form at all, by using nonparametric regression analysis. The nonparametric regression results are then used for appropriate functional specification of these functions g_i. Our data set, however, does not allow to take price effects into account. Recall, however, that price information was necessary to avoid identification problems. In our approach identification problems do not occur, as (in the first instance) no parametric functional form is specified.

The household expenditure functions we shall work with relate expenditures of household j on a certain group of commodities to net income (including children's allowance) of household j, the number of children in the age group 0-15 and the number of children in the age group 16 or over in household j. The latter only concerns children living with their parents

and having no income themselves.

In the econometric literature specific household expenditures are usually related to total expenditure, in order to impose the usual budget restriction and to interprete the model in terms of utility theory. A disadvantage of this approach is that the impact of demographic factors on total expenditure is ignored. It is conceivable that this impact is important, i.e., large households may spend a much larger fraction of their income on consumption (and thus save a much lower fraction of their income) than small households. By working with net income we therefore also take the effect of demographic factors on saving (or borrowing) into account, which gives a completer picture of the actual direct cost of children.

Since this study merely aims to be a pilot study of the applicability of nonparametric regression analysis in the empirical area under review, we keep the analysis here as simple as possible by distinguishing only two expenditure categories, namely

y_{1j} = expenditures of food, clothing and foot-wear,
y_{2j} = other expenditures

of household j. For the very same reason we only distinguish two age groups. The explanatory variables are now:

x_{1j} = net income,
x_{2j} = number of children in the age group 0-15,
x_{3j} = number of children in the age group 16 or over

of household j. The expenditure functions involved are:

$$y_{1j} = g_1(x_{1j}, x_{2j}, x_{3j}) + u_{1j}, \quad y_{2j} = g_2(x_{1j}, x_{2j}, x_{3j}) + u_{2j},$$

where the response functions (or regression functions) g_1 and g_2 are completely unknown, apart from the condition that g_1 and g_2 are continuously differentiable in x_{1j}. The disturbance terms u_{1j} and u_{2j} satisfy the usual condition that their conditional expectations relative to the regressors x_{1j}, x_{2j} and x_{3j} equal zero with probability 1:

$$E[u_{1j}|x_{1j},x_{2j},x_{3j}] = 0 \text{ and } E[u_{2j}|x_{1j},x_{2j},x_{3j}] = 0 \text{ with prob. 1.}$$

These conditions are no restrictions at all. They simply define the response functions g_1 and g_2 as conditional expectation functions, i.e.,

$$E[y_{1j}|x_{1j},x_{2j},x_{3j}] = g_1(x_{1j},x_{2j},x_{3j}) \text{ with prob. 1,}$$
$$E[y_{2j}|x_{1j},x_{2j},x_{3j}] = g_2(x_{1j},x_{2j},x_{3j}) \text{ with prob. 1.}$$

Note that these functions g_1 and g_2 are unique (with prob. 1), given the i.i.d. data generating process, in the sense that if there exists other functions f_1 and f_2, respectively, with the above properties then

$$P[g_i(x_{1j},x_{2j},x_{3j}) = f_i(x_{1j},x_{2j},x_{3j})] = 1, \quad i=1,2.$$

Moreover, the existence of g_1 and g_2 is guaranteed by the following mild conditions:

$$E|y_{1j}| < \infty, \quad E|y_{2j}| < \infty$$

Cf. Chung (1974, Theorem 9.1.1). Of course the uniqueness of g_1 and g_2 only applies to cross-section data: the expenditure system will likely change over time due to changes in preferences and prices. Moreover, we recall that no assumptions about the functional form of g_1 and g_2 will be made. We only assume that the variable x_{1j}, net income, is continuously distributed and that $g_1(x_1,x_2,x_3)$ and $g_2(x_1,x_2,x_3)$ are for each pair (x_2,x_3) continuously differentiable in x_1.

The procedure we advocate is the following. First we estimate g_1 and g_2 by nonparametric regression. The basic principles of the nonparametric regression approach and the results will be discussed in Section 3.3. Then we specify a parametric functional form in accordance with the nonparametric regression results, and this parametric model is estimated and tested in the usual way. This is the topic of Section 3.4. In Section 3.5 we discuss the estimation results. Finally, in Section 4 we consider the technical aspects of the nonparametric regression approach used in this paper.

3.2 *The data*

The data set we work with is the 1980 Budget Survey held by the Dutch Central Bureau of Statistics. This survey consists of an independent sample of 2859 households. For technical reasons we have split this sample in two subsamples of sizes 2000 and 859, respectively. The smaller subsample has been used for experiments with the nonparametric regression method, in order the improve the fit. Cf. Section 4. The larger subsample has been used for the actual nonparametric and parametric estimation of our expenditure functions.

A typical feature of the budget survey involved is that total expenditures may exceed net income, especially in the low income range. This is due to the fact that expenditures on durables are completely attributed to the year of purchase. Thus, if a household buys say new furniture in a certain year, the total amount of the purchase involved is considered as an expenditure in that year, even if the purchase has been financed by a loan. The same applies to clothing and foot-wear: although a suit or a pair of shoes may last longer than a year the total amount of the purchase is considered as expenditures in the year of the purchase. As a consequence, adding up (i.e., $y_{1j} + y_{2j} = x_{1j}$) does not apply.

Since the 1980 Budget Survey is a representative survey, it also contains households with only one parent and households of elderly. These households have been excluded from our analyses (after splitting the sample in two subsamples). However, the remaining data subsets of sizes 1130 and 552, respectively, are then no longer random samples, a situation not accounted for in the theory of nonparametric estimation. As will be shown in Section 4 a simple modification of the nonparametric regression approach will correct for that.

Finally we note that the further subsample of size 1130 contains five households with expenditures on food, clothing and foot-wear exceeding net income, 86 households with other expenditures exceeding net income and 424 households with total expenditure exceeding net income. For the further subsample of size 552 these numbers are 1, 48 and 226, respectively. This is mainly due to the typical way expenditures are measured in the budget survey under review, although we do not exclude that also occasional mea-

surement errors in net income may contribute to this phenomenon (despite the assurance of CBS that in the survey under review income is accurately measured).

3.3 *Nonparametric regression: basic principles and results*

In this subsection we discuss in a non-technical manner the nonparametric regression approach and the nonparametric regression results for the expenditure functions under review. The technical details will be given in Section 4.

As said before, nonparametric regression is a statistical technique by which we can substract information from the data about the functional form of a regression model without restricting this functional form to a particular parametric family of functional forms. Given a sample $\{(y_1,x_1),\ldots,(y_n,x_n)\}$ from a k+1-variate distribution, where y_j is the dependent variable and x_j is a k-vector of regressors, the usual approach is to specify in advance a parametric family $f(x,\beta)$ of regression functions such that for some particular parameter value β_0,

$$E[y_j|x_j] = f(x_j,\beta_0) \text{ with prob. 1.}$$

The parametric regression model then takes the form

$$y_j = f(x_j,\beta_0) + u_j;\ E[u_j|x_j] = 0 \text{ with prob. 1.}$$

In practice the most popular specification of this parametric family of functional forms is the linear family:

$$f(x,\beta) = x'\beta \qquad \text{(without constant term)}$$

or

$$f(x,\beta) = (1,x')\beta \text{ (with constant term).}$$

Given the choice of the parametric family $f(x,\beta)$ of functional forms, the estimation of the model now merely amounts to estimation of the unknown parameter vector β_0.

In nonparametric regression analysis we do not assume a parametric family of functional forms. The response function g(x) of the regression model

$$y_j = g(x_j) + u_j;\ E[u_j|x_j] = 0 \text{ (with prob. 1)}$$

is completely unknown and has to be estimated entirely from the data. There are various techniques to do that. Here we have used the *kernel regression approach*. The basic idea of kernel regression analysis is to form a weighted sum of the y_j's, where the weights depend on the distance between x_j and some fixed x. There is quite a variety of admissible choices for this weight function, but this need not concern us at the present stage. What is important now is to know that it is possible to specify a sequence $\{W_{nj}(.,.)\}$ of weight functions depending on the sample size n and the observation index j such that the random function

$$\hat{g}(x) = \sum_{j=1}^{n} y_j W_{nj}(x,x_j)$$

is a consistent estimator of the unknown response function g(x). One may consider this weight function as a sort of inverse measure of the distance between x and x_j, i.e., the closer x_j is to x, the larger $W_{nj}(x,x_j)$ will be and thus the more weight is put on the corresponding y_j. As we shall see in Section 4, the construction of the weight function $W_{nj}(.,.)$ does not involve any explicit knowledge of the true response function g(x).

In order to illustrate the basic idea behind nonparametric regression, assume for the moment that the regressors x_j are discretely distributed. In particular, assume that x_j takes with probability 1 values in a finite set X. Moreover, let $x \in X$. Specifying $W_{nj}(.,.)$ such that

$$W_{nj}(x,x_j) = I(x_j=x) \ / \ \sum_{\ell=1}^{n} I(x_\ell = x),$$

where I(.) is the indicator function, i.e.

$$I(x_j=x) = 1 \text{ if } x = x_j,$$
$$= 0 \text{ if } x \neq x_j,$$

the estimator $\hat{g}(x)$ is then just the mean of the y_j's corresponding to the x_j's equal to x. It is well-known that in this case $\hat{g}(x)$ is a consistent estimator of $g(x)$. Thus in the purely discrete case nonparametric regression amounts to classifying the y_j's into a number of cells, each corresponding to one of the possible outcomes of x_j, and to use the mean of the y_j's in each cell as an estimate of the conditional expectation of y_j relative to the event that x_j belongs to the cell involved.

In general regression problems where some or all of the regressors are continuously distributed things are not so simple as above. Nevertheless also then it is possible to specify suitable weight functions such that $\hat{g}(x)$ is pointwise or uniformly consistent and even asymptotically normally distributed. The latter result is of the form

$$r_n[\hat{g}(x) - g(x)] \to N[0,\sigma_g^2(x)] \quad \text{in distr.},$$

where the rate of convergence r_n satisfies $r_n \to \infty$ as $n \to \infty$. More generally, for distinct non-random points $x^{(1)}, x^{(2)},\ldots,x^{(M)}$ we have

$$r_n \begin{pmatrix} \hat{g}(x^{(1)}) - g(x^{(1)}) \\ \vdots \\ \hat{g}(x^{(M)}) - g(x^{(M)}) \end{pmatrix} \to N_M[0,\Sigma_M] \text{ in distr.}$$

where Σ_M is a *diagonal* matrix with diagonal elements

$$\sigma_g^2(x^{(1)}),\ldots,\sigma_g^2(x^{(M)}).$$

Thus the random variables

$$r_n[\hat{g}(x^{(1)})-g(x^{(1)})],\ldots,r_n[\hat{g}(x^{(M)})-g(x^{(M)})]$$

are asymptotically *independent*.

The rate of convergence in distribution, r_n, depends on the nature of the distribution of the components of x_j (discrete versus continuous, or

mixed) and on the specification of the weight functions $W_{nj}(.,.)$. In the case under review where $x_j = (x_{1j}, x_{2j}, x_{3j})'$ with x_{1j} continuously distributed and x_{2j} and x_{3j} discretely distributed this rate is $r_n = n^{8/17}$ (cf. Section 4), which is slightly slower than the rate $\sqrt{n}$ which applies to the convergence in distribution of parameter estimators.

It is possible to construct a consistent estimator $\hat{\sigma}_g^2(x)$ of the variance $\sigma_g^2(x)$. This allows us to construct 95% confidence intervals on the basis of the result

$$r_n[\hat{g}(x) - g(x)]/\hat{\sigma}_g(x) \rightarrow N(0,1) \text{ in distr.},$$

namely the interval

$$[\hat{g}(x) - 1.96\hat{\sigma}_g(x)/r_n \ , \ \hat{g}(x) + 1.96\hat{\sigma}_g(x)/r_n] \ .$$

The nonparametric regression results for the expenditure functions under review are displayed in Figures 1 to 16. The first 8 figures show the kernel regression estimator $\hat{g}(x_1, x_2, x_3)$ (the solid line) for expenditures on food, clothing and foot-wear, where x_1 (net income) runs from 16,000 to 65,000 guilders. In some cases the income range is smaller, due to lack of observations in the low and high income range. The scale of the figures is linear on both axes. Each figure corresponds to a household type (x_2, x_3), where x_2 is the number of children in the age group 0-15 and x_3 is the number of children in the age group 16 or over. We only show the nonparametric results for households with $0 \leq x_2 \leq 3$ and $0 \leq x_3 \leq 1$, as other households are too rare. The dotted lines are the 95% confidence bands. Observe that the 95% confidence band becomes wider in the low and high income range, due to lack of observations. The other 8 figures show the nonparametric results for other expenditures. In the next section we shall interprete these nonparametric regression results. In particular we shall pay attention to the question how to translate these results to parametric specifications.

Figure 1

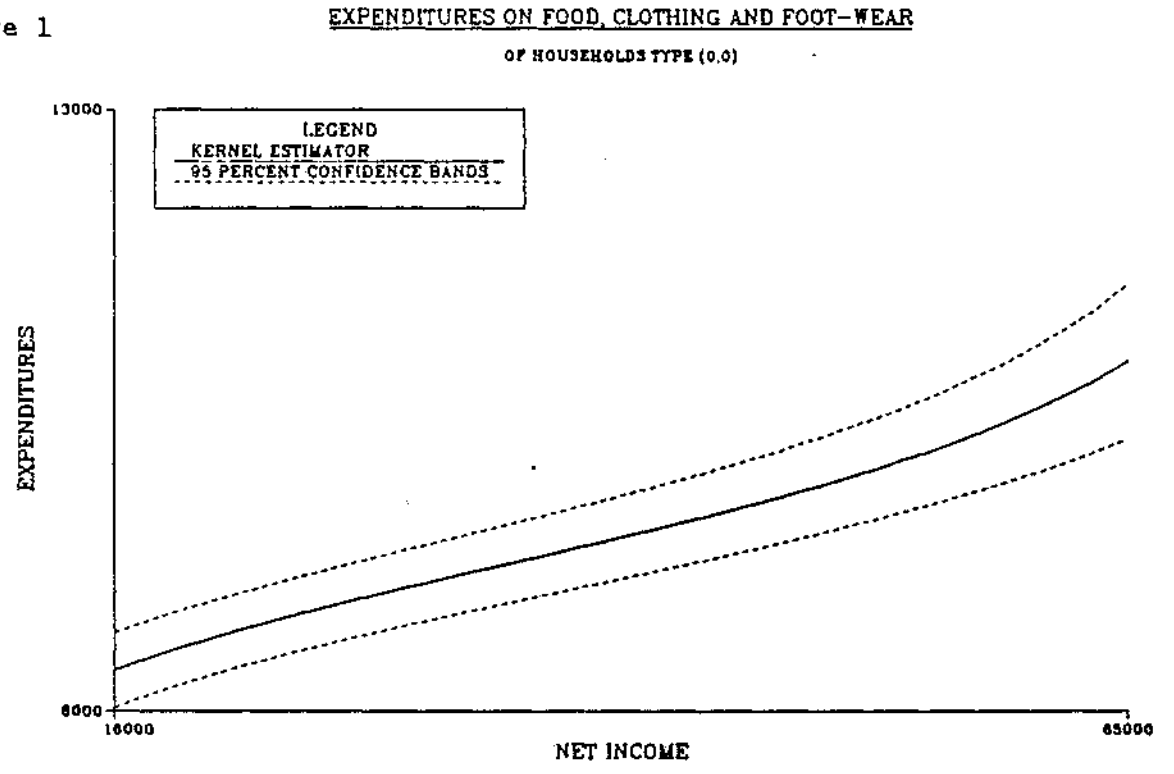


Figure 2

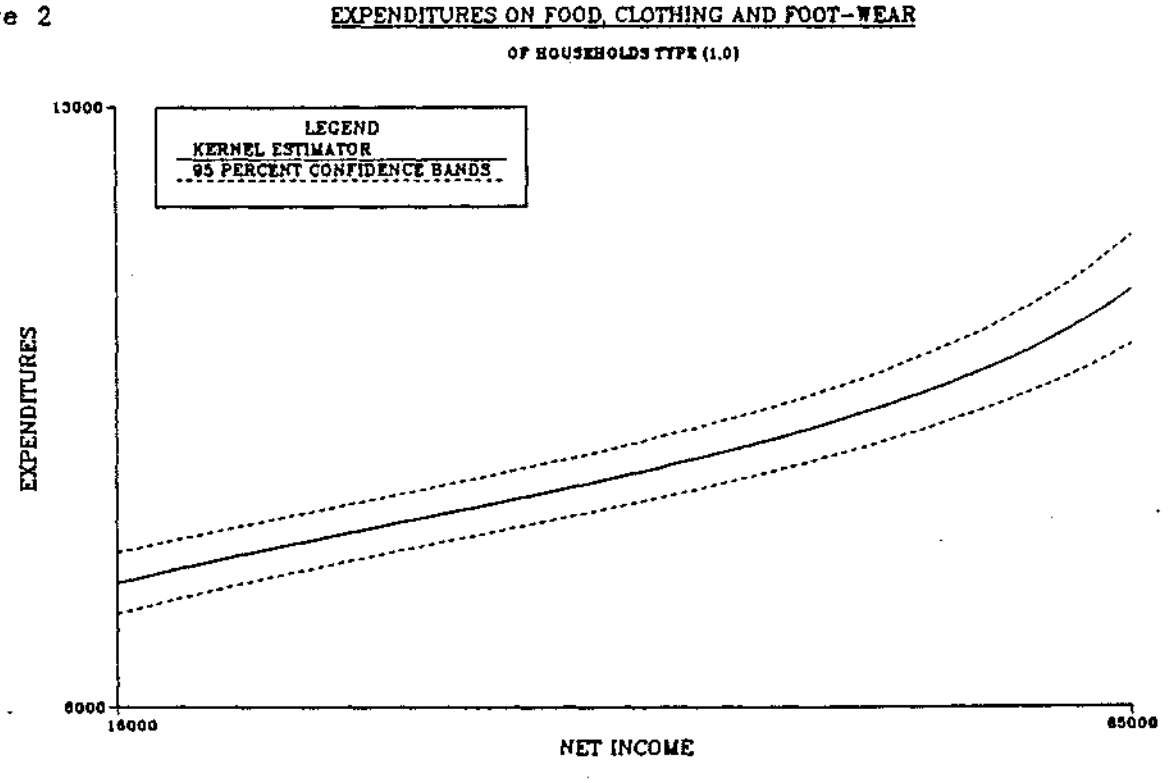


Figure 3

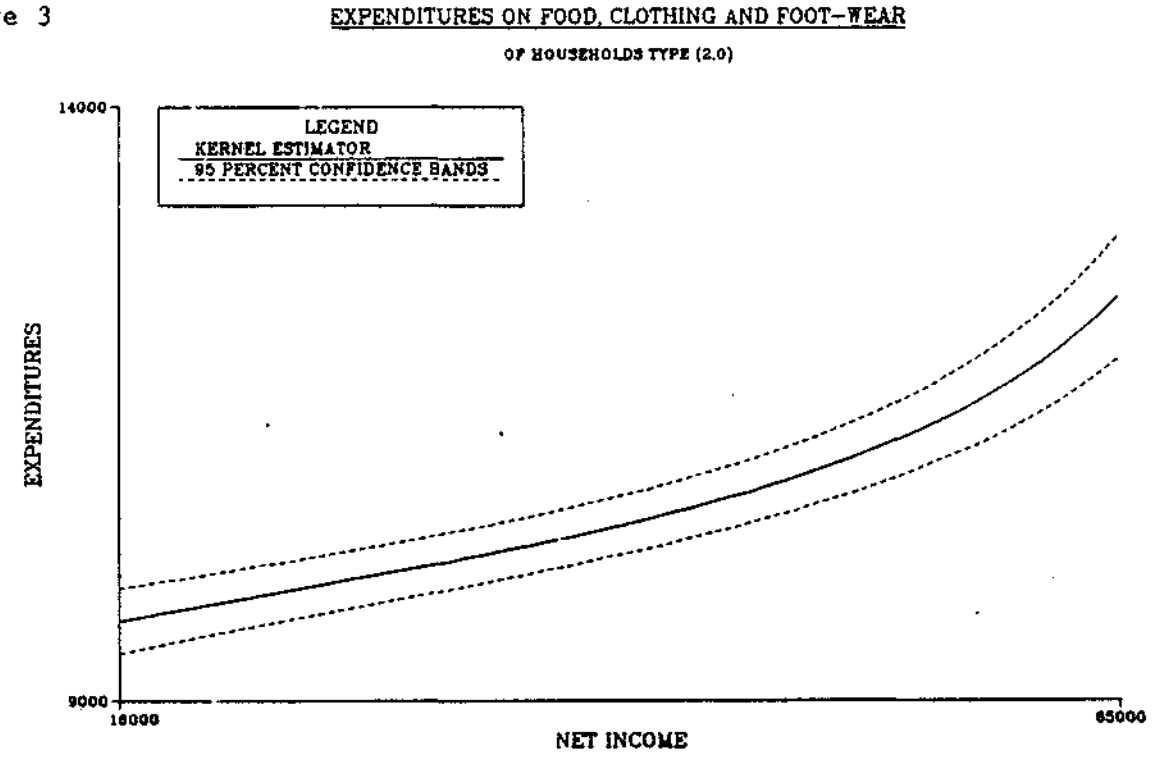


Figure 4

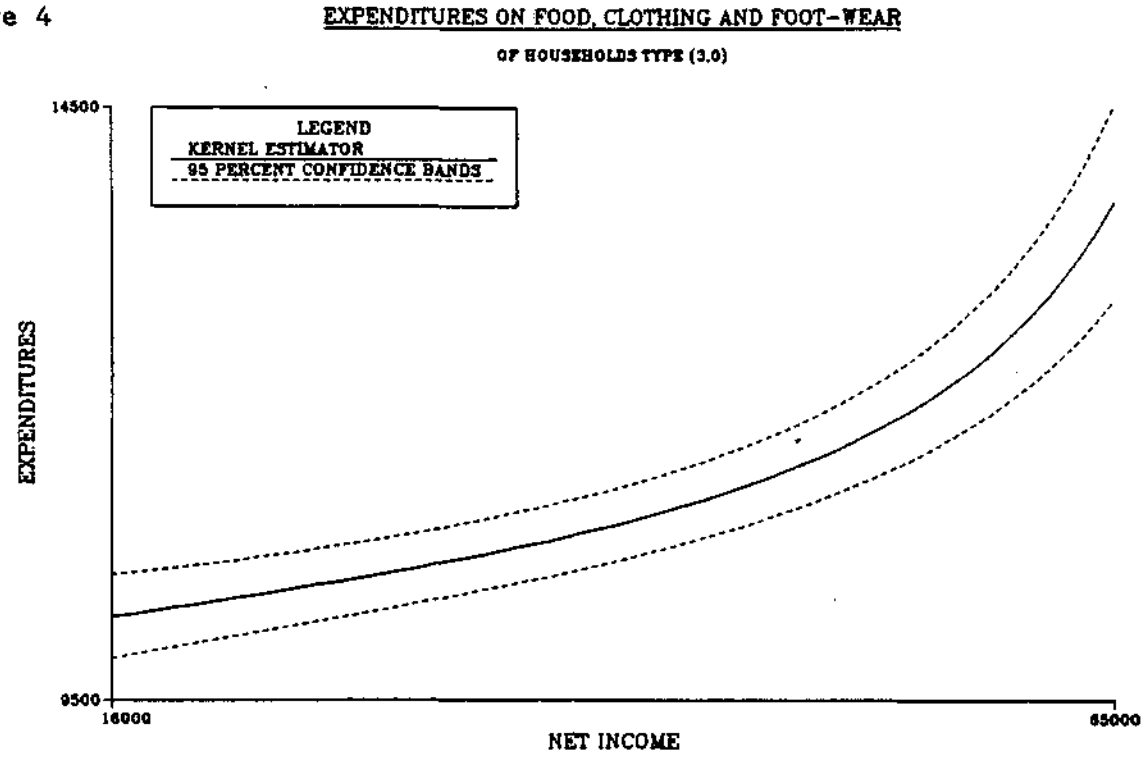


Figure 5

EXPENDITURES ON FOOD, CLOTHING AND FOOT-WEAR

OF HOUSEHOLDS TYPE (0.1)

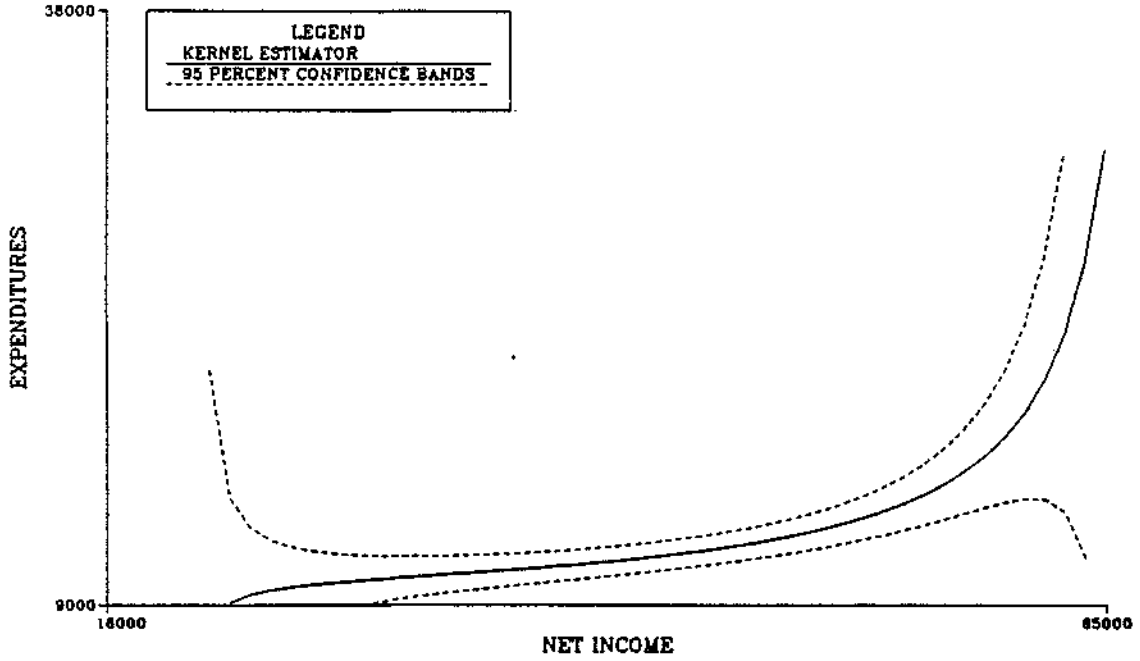


Figure 6

EXPENDITURES ON FOOD, CLOTHING AND FOOT-WEAR

OF HOUSEHOLDS TYPE (1.1)

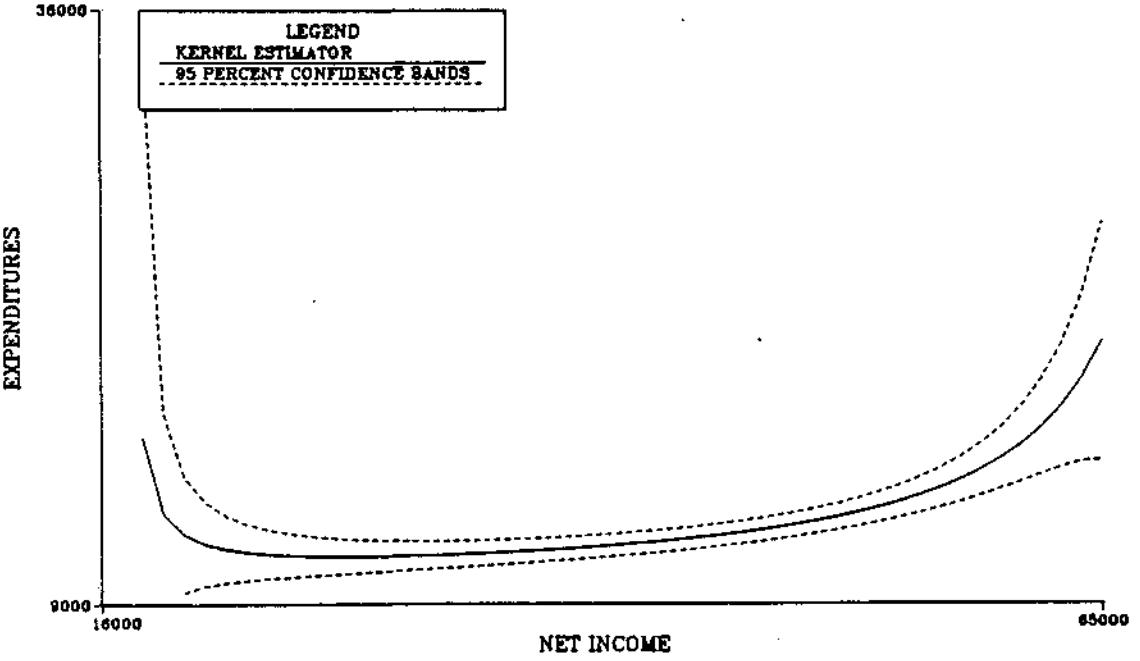


Figure 7

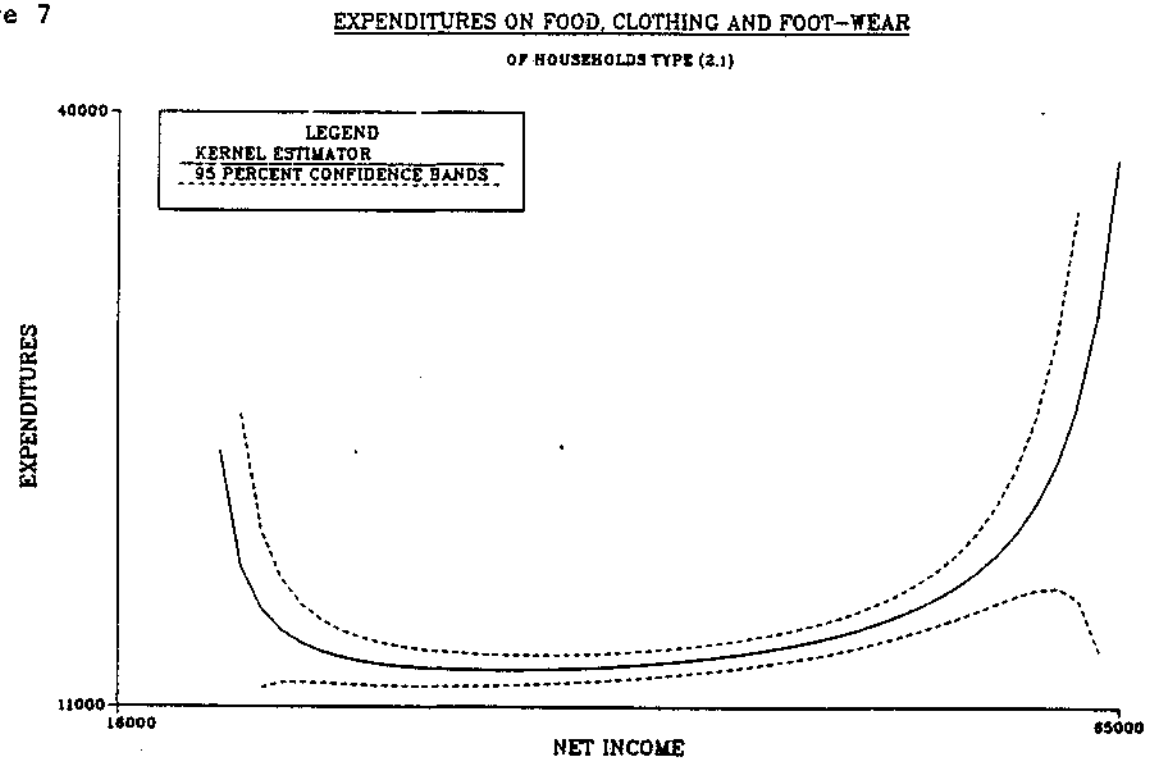


Figure 8

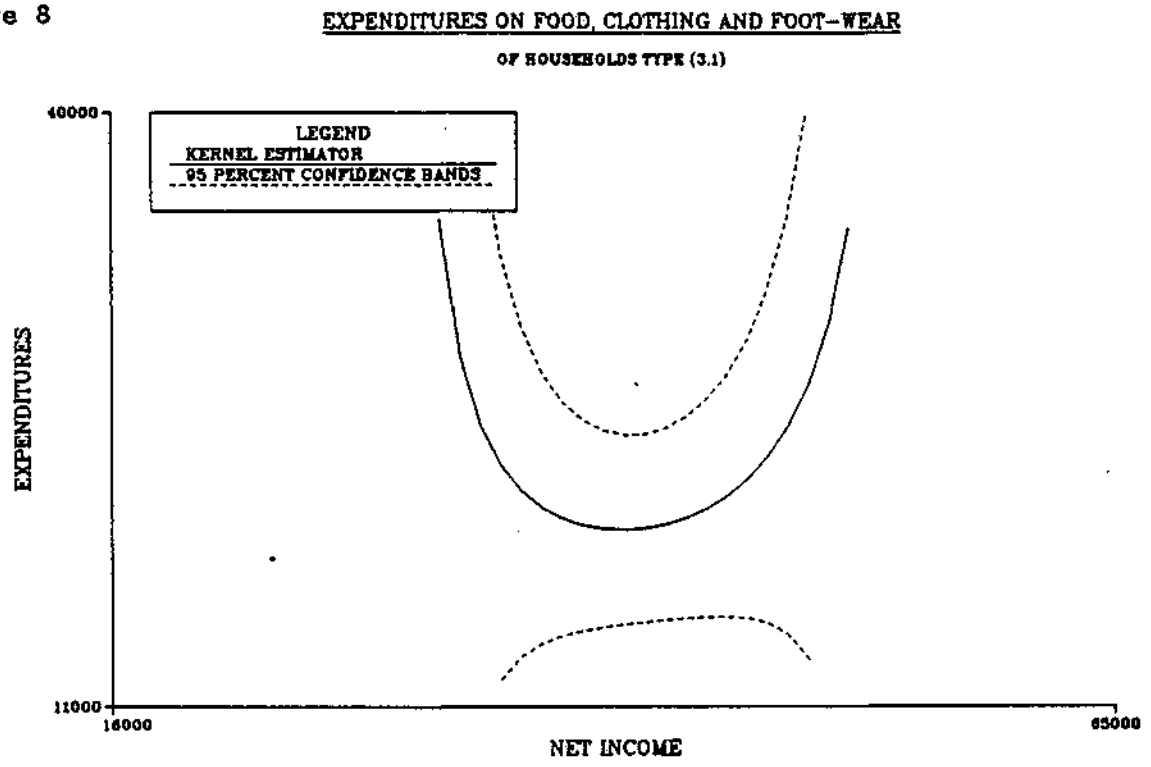


Figure 9 OTHER EXPENDITURES OF HOUSEHOLD TYPE (0,0)

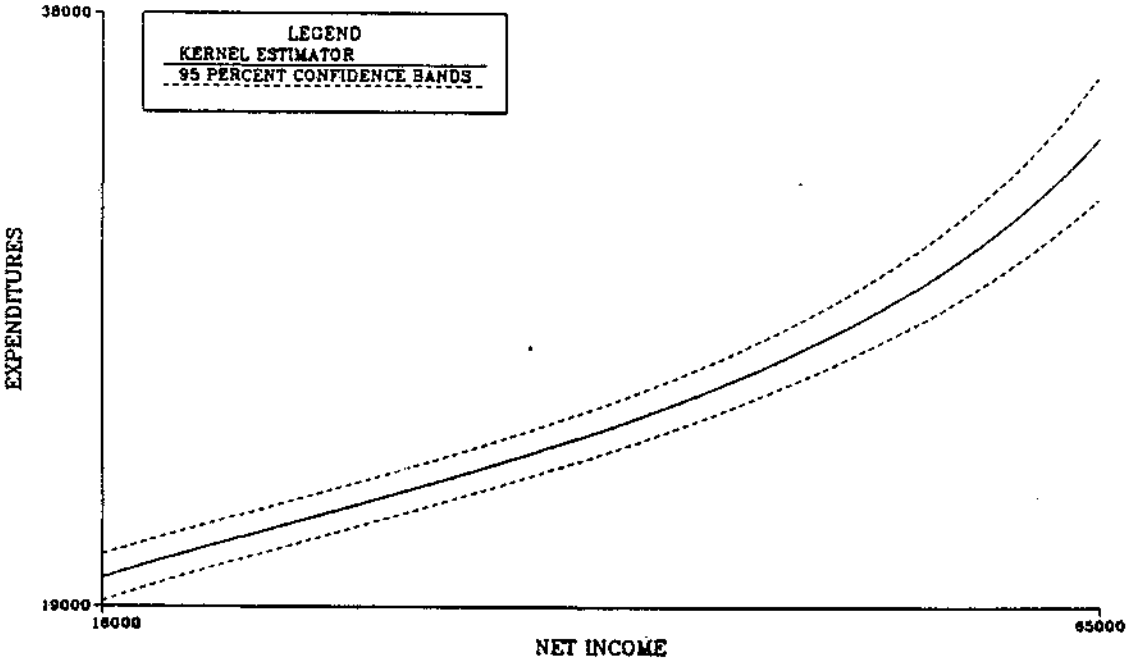


Figure 10 OTHER EXPENDITURES OF HOUSEHOLD TYPE (1,0)

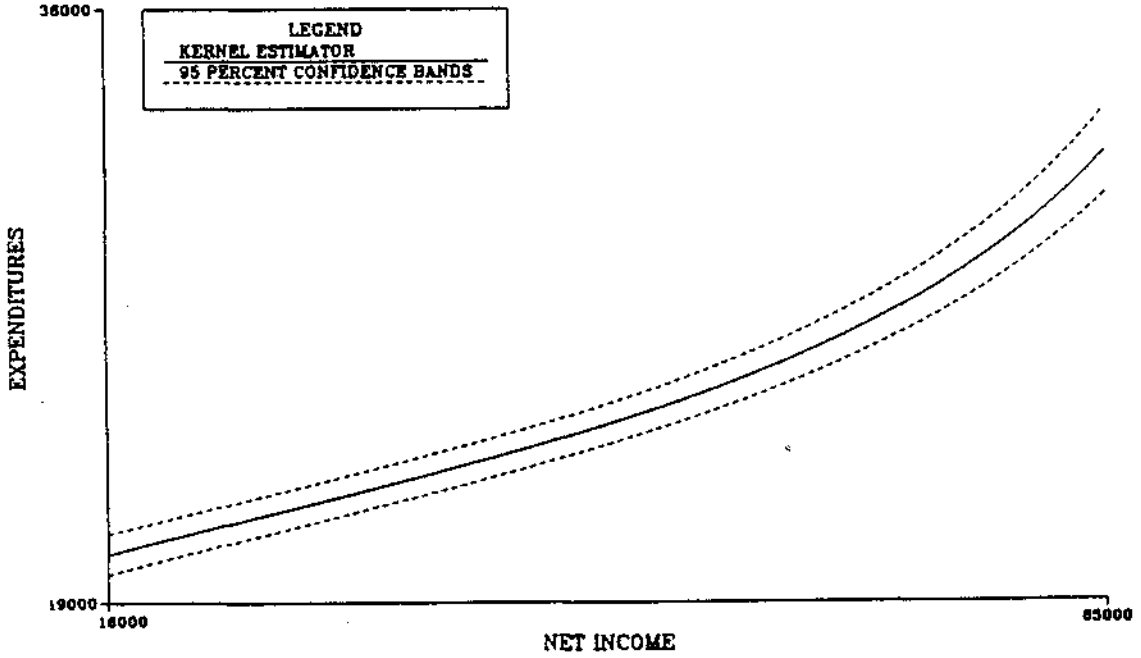


Figure 11 OTHER EXPENDITURES OF HOUSEHOLD TYPE (2,0)

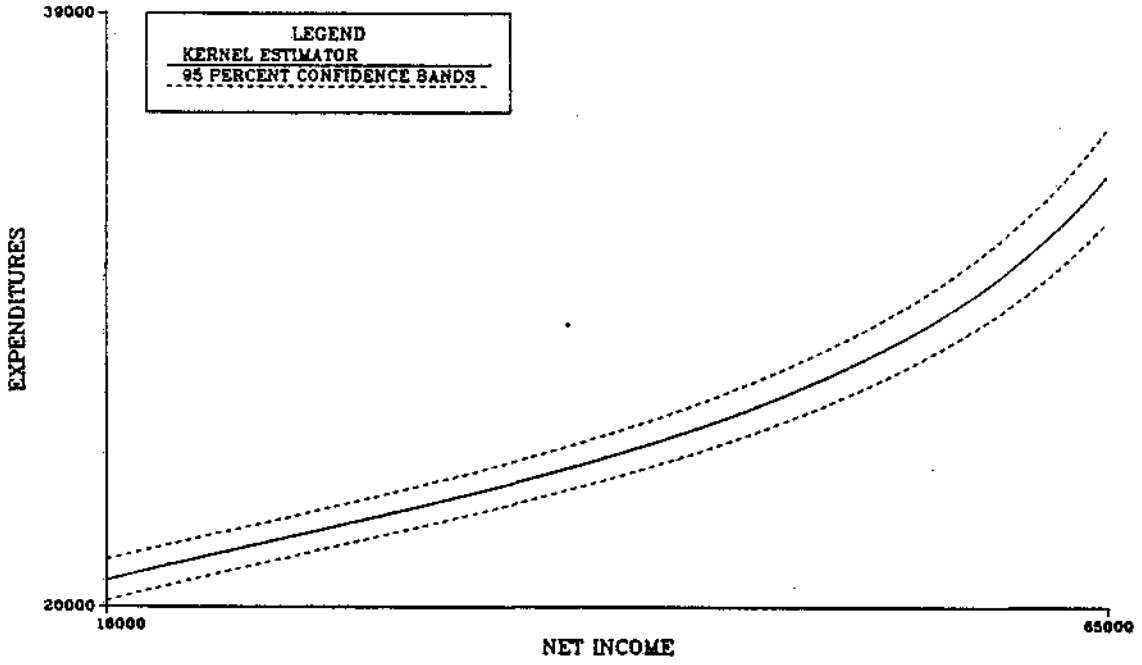


Figure 12 OTHER EXPENDITURES OF HOUSEHOLD TYPE (3,0)

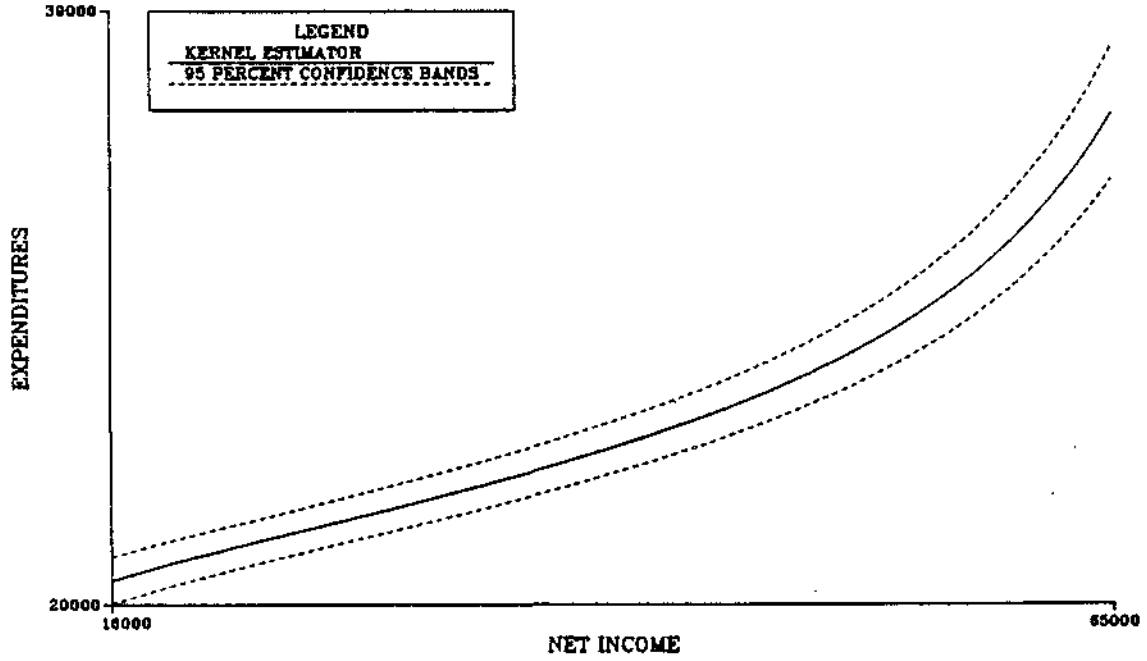


Figure 13

OTHER EXPENDITURES OF HOUSEHOLD TYPE (0,1)

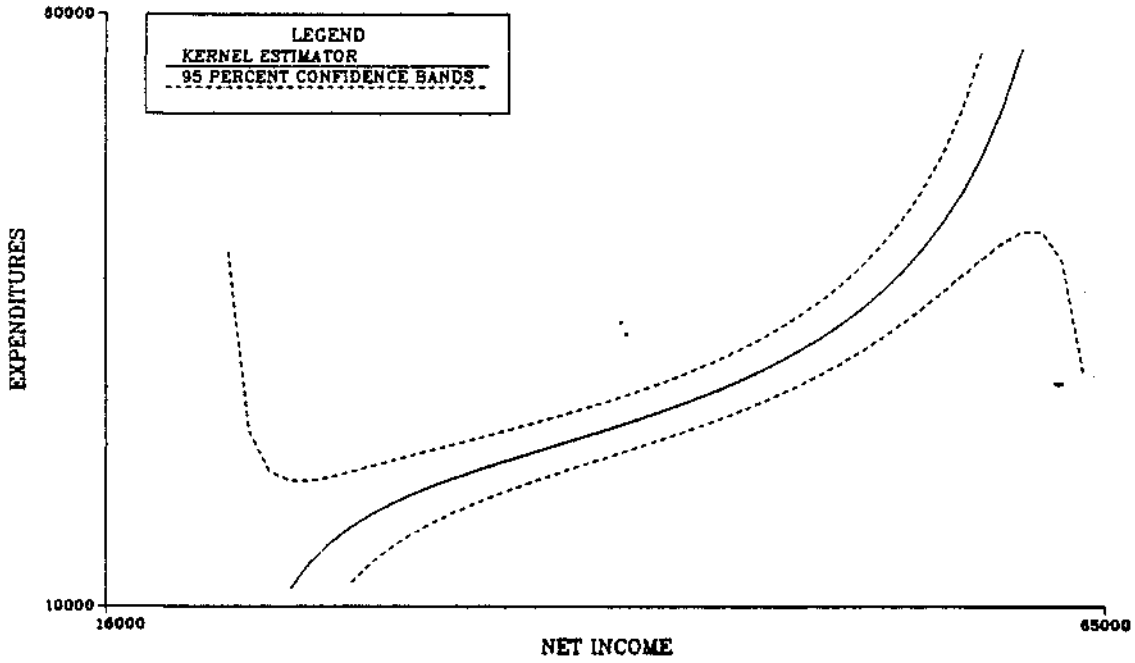


Figure 14

OTHER EXPENDITURES OF HOUSEHOLD TYPE (1,1)

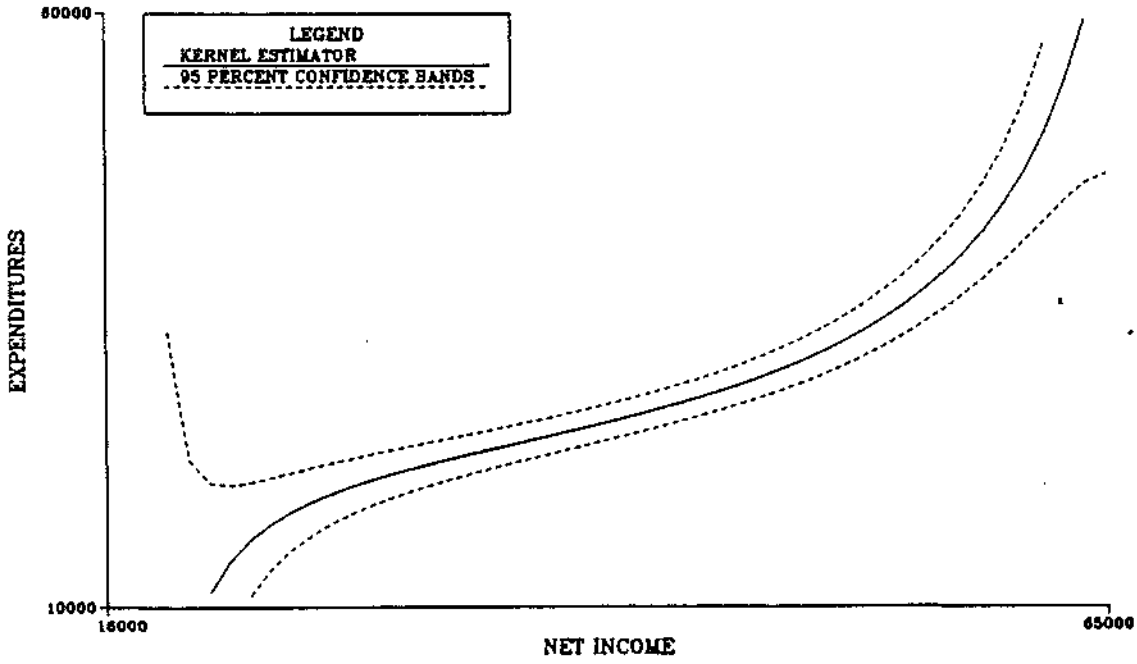


Figure 15 OTHER EXPENDITURES OF HOUSEHOLD TYPE (2,1)

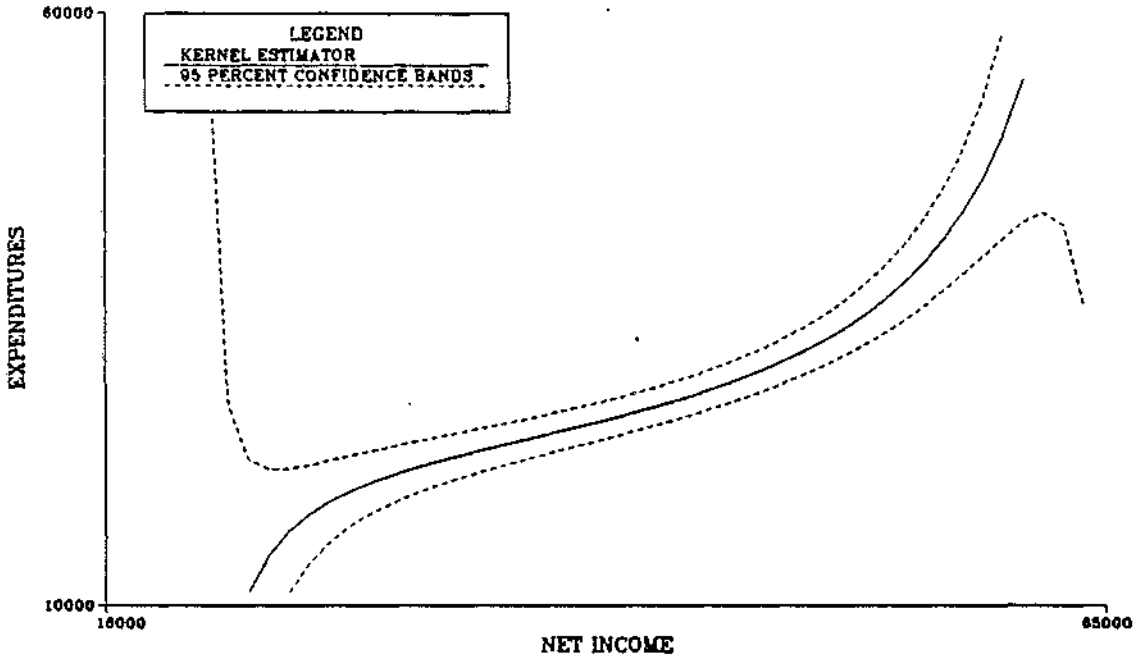
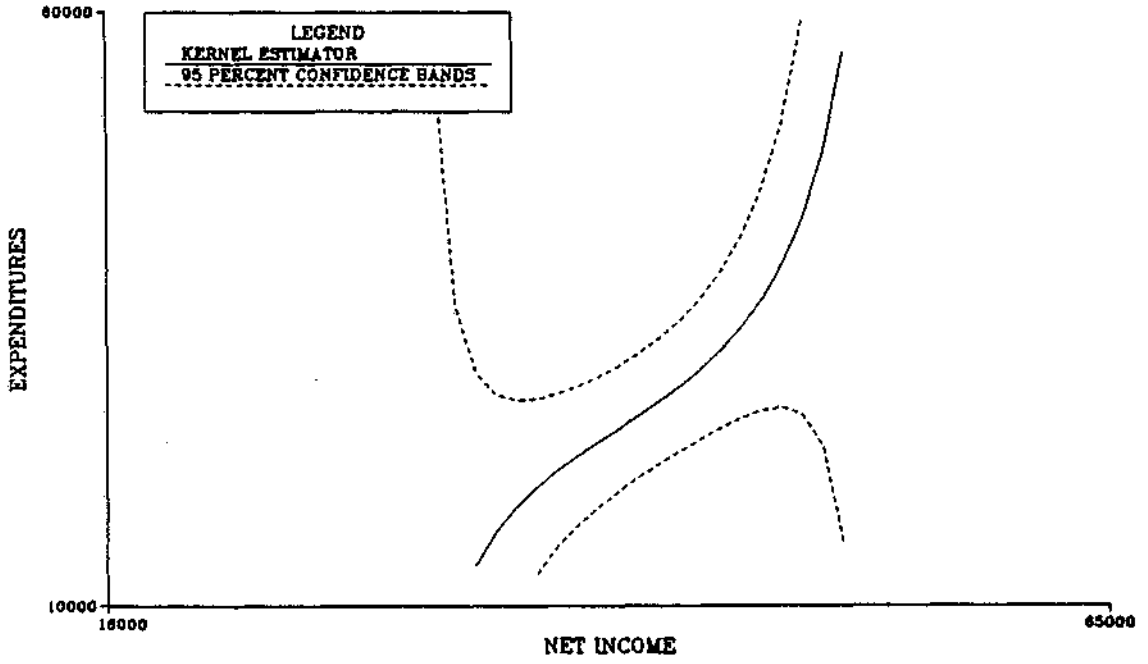


Figure 16 OTHER EXPENDITURES OF HOUSEHOLD TYPE (3,1)



3.4 *Parametrisation of the nonparametric regression results*

The nonparametric regression results for expenditures on food, clothing and foot-wear indicate that the income-expenditure relationships involved are almost linear: for each household type it is possible to draw a straight line almost entirely inside the 95% confidence band. The estimated Engel curves are only bending in the low and high income ranges. These curved parts, however, need not be significantly different from a straight line, as indicated by the 95% confidence bands, as in nonparametric regression analyses estimation errors manifest themselves in the form of bumps on the estimated regression curve. Thus the nonparametric regression results indicate that over the income range 16,000-65,000 the Engel curves involved are linear. The same applies to other expenditures. Nevertheless we have specified these Engel curves as third-order polynomials in net income, in order to catch the bending parts as well and to test whether the actual Engel curves are indeed linear.

In order to check this specification we have approximated the kernel regression function estimators for each household type by third-order polynomials in net income, by regressing the kernel estimator on x_1, x_1^2 and x_1^3 for grid points in the interval 16,000-65,000. It appears that for each household type this polynomial approximation fits in the 95% confidence band, which indicates that a third-order polynomial is a suitable functional form for the expenditure functions under review. The third-order polynomial approximations are in fact so close that they hardly can be distinguished from the corresponding kernel estimators on the income range 16,000-65,000. Therefore we cannot show them in the figures.

The parameters of the third-order polynomials can be made dependent of the number of children in the household by using the following dummy variables.

$$d_{mj} = 1 \text{ if } x_{2j} = m,$$
$$= 0 \text{ if } x_{2j} \neq m, \; m=1,2,3,$$
$$d_{4j} = x_{3j} \quad (x_{3j} \leq 1).$$

The parametrisation of the nonparametric regression results then becomes:

$$y_{ij} = \alpha_{i0} + \beta_{i0}x_{1j} + \delta_{i0}x_{1j}^2 + \varepsilon_{i0}x_{1j}^3 + \sum_{m=1}^{4}(\alpha_{im}d_{mj} + \beta_{im}d_{mj}x_{1j} + \delta_{im}d_{mj}x_{1j}^2 + \varepsilon_{im}d_{mj}x_{1j}^3) + u_{ij}, \quad i=1,2.$$

We have used a further subsample of the subsample of size 2000 to estimate the parameters involved. This further subsample consists of all households of the type (x_2, x_3) with $x_2 \leq 3$ and $x_3 \leq 1$, net income x_1 in the range 16,000-65,000, and two parents both younger than 65. The size of this further subsample is 1010. The OLS results are given in Table 1.

The test of the linearity hypothesis amounts to testing the null hypo-

Table 1. OLS results for the third-order polynomial

	Food, clothing & foot-wear		Other expenditures	
	OLS estimates	t-values	OLS estimates	t-values
α_0	1508.	0.2592	19590.	1.276
α_1	4639.	0.5478	-613.6	-0.0255
α_2	5169.	0.6001	-22330.	-1.087
α_3	10440.	0.6694	12350.	0.3290
α_4	-8817.	-0.9187	-18630.	-0.6248
β_0	0.3659	0.7723	-0.7216	-0.5542
β_1	-0.3715	-0.5457	0.1205	0.06028
β_2	-0.2454	-0.3528	1.629	0.9433
β_3	-0.5299	-0.4449	-0.6945	-0.2387
β_4	0.9577	1.226	2.122	0.8807
δ_0	-0.5506E-5	-0.4455	0.3428E-4	0.9759
δ_1	0.1102E-4	0.6295	-0.2394E-5	-0.0455
δ_2	0.4840E-5	0.2686	-0.3800E-4	-0.8217
δ_3	0.1065E-4	0.3647	0.9124E-5	0.1254
δ_4	-0.2532E-4	-1.261	-0.6889E-4	-1.119
ε_0	0.3458E-10	0.3319	-0.3055E-9	-1.007
ε_1	-0.1030E-9	-0.7176	0.1177E-10	0.0268
ε_2	-0.2521E-10	-0.1683	0.3025E-9	0.7671
ε_3	-0.6848E-10	-0.2961	-0.1315E-11	-0.0022
ε_4	0.2066E-9	1.264	0.6634E-9	1.319
R^2	0.1722		0.3227	
SE	3354.		8099.	

thesis

$$H_0: \delta_{im}=\varepsilon_{im}=0 \text{ for } m=0,1,2,3,4.$$

The test statistics of the Wald test involved are 6.394 for expenditures on food, clothing and foot-wear and 7.820 for other expenditures. Under the null hypothesis these test statistics are asymptotically χ^2_{10} distributed, hence the linearity hypothesis cannot be rejected at any reasonable significance level.

Next we have tested whether the linearity hypothesis holds with *constant slope*. Thus the null hypothesis to be tested is now:

$$H_0: \beta_{im}=0 \text{ for } m=1,2,3,4;\ \delta_{im}=\varepsilon_{im}=0 \text{ for } m=0,1,2,3,4.$$

The test statistics of the Wald test are 9.568 for expenditures on food, clothing and foot-wear and 13.23 for other expenditures. Under H_0 these test statistics are asymptotically χ^2_{14} distributed, and consequently also this null hypothesis cannot be rejected. Thus the model reduces to:

$$y_{ij} = \alpha_{i0} + \alpha_{i1}d_{1j} + \alpha_{i2}d_{2j} + \alpha_{i3}d_{3j} + \alpha_{i4}d_{4j} + \beta_{i0}x_{1j} + u_{ij},\ i=1,2.$$

Furthermore, we have tested whether this model can be written as a linear regression model with explanatory variables x_{1j}, x_{2j} and x_{3j}. This simplification corresponds to the following hypothesis:

$$H_0: \alpha_{i2}=2\alpha_{i1};\ \alpha_{i3}=3\alpha_{i1};\ \beta_{im}=0 \text{ for } m=1,2,3,4;$$

$$\delta_{im}=\varepsilon_{im}=0 \text{ for } m=0,1,2,3,4.$$

The Wald statistics involved are 10.97 for food, clothing and foot-wear and 16.15 for other expenditures. Under the null these statistics are asymptotically χ^2_{18} distributed and therefore we cannot reject the null hypothesis.

Finally we have tested whether the Engel curves are linear and independent of the household size. This hypothesis corresponds to:

H_0: $\alpha_{im}=\beta_{im}=0$ for m=1,2,3,4; $\delta_{im}=\varepsilon_{im}=0$ for m=0,1,2,3,4.

The Wald statistics are 119.5 for food, clothing and foot-wear and 16.33 for other expenditures. Since $P[\chi^2_{18} > 119.5]$ cannot be distinguished from zero we have to reject this null hypothesis for expenditures on food, clothing and food-wear, while the hypothesis involved cannot be rejected for other expenditures.

Table 2 summarizes the test results. Note that the tests involved are not independent. From a formal point of view we should therefore not re-estimate the model after each test as otherwise the type I errors may accumulate. Nonetheless we have checked the final conclusions by conducting a similar sequence of tests starting from the linear model with slope and intercept depending on the family size, and the linear model with constant slope and intercept depending on family size, respectively. These tests lead to the same conclusion as before, namely that the expenditure function for expenditures on food, clothing and foot-wear is a linear function in net income (x_1), the number of children in the age group 0-15 (x_2) and the number of children in the age group 16 or over (x_3), while the expenditure

Table 2: Test results

H_0:	Degr. of free-dom (=ℓ)	Food, clothing & foot-wear		Other expenditures	
		Wald stat. (=W)	$P[\chi_\ell^2 > W]$	Wald stat. (=W)	$P[\chi_\ell^2 > W]$
$\delta_{im}=\varepsilon_{im}=0, m=0,..,4$	10	6.394	0.78	7.820	0.65
$\beta_{im}=0$, m=1,2,3,4 $\delta_{im}=\varepsilon_{im}=0, m=0,..,4$	14	9.568	0.79	13.23	0.51
$\alpha_{i2}=2\alpha_{i1}, \alpha_{i3}=3\alpha_{i1}$, $\beta_{im}=0$, m=1,2,3,4 $\delta_{im}=\varepsilon_{im}=0, m=0,..,4$	16	10.97	0.81	16.15	0.44
$\alpha_{im}=\beta_{im}=0, m=1,..,4$ $\delta_{im}=\varepsilon_{im}=0, m=0,..,4$	18	119.5	0.0	16.33	0.57

function for other expenditures is a linear function in net income only.

3.5 *Conclusions*

The simplification of the polynomial model suggested by the test results in Section 3.4 now lead to the following models:

Food, clothing and foot-wear:

$$\hat{y}_1 = 5407. + 0.09937\ x_1 + 775.1\ x_2 + 2106.\ x_3 \qquad (3.1)$$
$$(14.5) \quad (9.868) \quad (7.185) \quad (6.819)$$
$$R^2 = 0.1667;\ SE = 3338.$$

Other expenditures:

$$\hat{y}_2 = 5671. + 0.5163\ x_1 \qquad (3.2)$$
$$(5.921) \quad (19.11)$$
$$R^2 = 0.3110;\ SE = 8095.$$

Note that model (3.1) can be written as

$$\hat{y}_1/m(x_2, x_3) = 2703.5 + 0.09937\ x_1/m(x_2, x_3) \qquad (3.3)$$

where

$$m(x_2, x_3) = 2 + 0.2867\ x_2 + 0.7790\ x_3 \qquad (3.4)$$

is the adult equivalence scale. It should be noted, however, that in the literature equivalence scales are usually derived from expenditure systems relating expenditures on groups of commodities to *total expenditure* rather than income. The above equivalence scale is therefore not quite compatible with the equivalence scales found in the literature, although its interpretation is the same. Thus, as far as food, clothing and foot-wear is concerned a child under 16 counts for 28,67% of an adult and a dependent child of 16 or over counts for 77.9% of an adult, in a household with two

parents and net income in the range 16,000-65,000. Moreover, an additional child under 16 induces additional expenditures on food, clothing and footwear to the amount of about 775 guilders per year, whereas an additional child of 16 or over induces an additional amount of 2106 guilders.

It should be stressed that the lack of impact of household size on the other expenditures does not imply that there is no impact at all. It is likely that the extra expenditures due to children will be covered by substitution within the same expenditure category. For example the extra expenditures on housing may be counterbalanced by cheaper vacations, a second hand car rather than a new one, etc.

The subsample of size 1010 on which the final estimation results were based contains one household with $y_1 > x_1$, 73 households with $y_2 > x_1$ and 380 households with $y_1+y_2 > x_1$. The model indicates that the latter occurs if $x_1 < 28,824 + 2,017.\ x_2 + 5,480.\ x_3$.

As said before, the expenditure functions considered in the literature usually relate expenditures on various commodities to total expenditure rather than to income, in order to impose the budget contraint and to interprete the models in terms of utility theory. We can put the above models in this form by solving equation (3.2) to x_1 and substituting the result for x_1 in equation (3.2). This yields, after some elementary calculations:

$$\hat{y}_1 = 3619 + 0.1614\ \hat{y} + 650\ x_2 + 1766\ x_3 \qquad (3.5)$$

where $\hat{y} = \hat{y}_1 + \hat{y}_2$. This model can also be written as

$$\hat{y}_1/m(x_2,x_3) = 1809.5 + 0.1614\ \hat{y}/m(x_2,x_3) \qquad (3.6)$$

with

$$m(x_2,x_3) = 2 + 0.3592\ x_2 + 0.976\ x_3 \qquad (3.7)$$

the corresponding equivalence scale.

It should be noted that model (3.5) relates $E(y_1|x_1,x_2,x_3)$ to $E(y_1+y_2|x_1,x_2,x_3)$ rather than y_1 to y_1+y_2. The equivalence scale (3.7) is therefore still not fully compatible with the scales found in the liter-

ature, although more compatible than (3.4). From (3.7) it follows that, as far as expenditures on food, clothing and foot-wear are concerned, a child under 16 counts for about 36% of an adult, and a child of 16 or over counts for 98% of an adult.

One may object against our approach and our results that there is no relationship at all with economic theory (in particular utility theory) and that the above results cannot be interpreted in terms of utility theory. Indeed, we actually have worked the other way around, i.e., we started with analysing the data in order to determine a model rather than setting up first the model in order to analyse the data. One should bear in mind, however, that the classical econometric approach reviewed in Section 2 assumes very restrictive household behavior, in particular the implicit assumption that all households are faced with exactly the same utility function. This will unlikely be the case in reality. Our concern merely is to determine actual household behavior, regardless whether or not this behavior is rational. Furthermore, to the best of our knowledge none of the authors of the econometric papers mentioned in Section 2 have properly tested the functional form of their models against misspecification, so their conclusions regarding the cost of children might be biased.

4. THE KERNEL REGRESSION APPROACH

4.1. *The modified kernel regression estimator for the i.i.d. mixed continuous-discrete case*

In this section we summarize Bierens' (1987) modified kernel regression function estimation approach for the case of mixed continuous-discrete expanatory variables and an i.i.d. data generating process. We start with the description of the data generating process.

Assumption 1. Let $(y_1,x_1),..,(y_n,x_n)$ be i.i.d. random vectors, where the y_j's are the dependent variables and the x_j's are k-component vectors of regressors. Moreover,

$$E|y_j|^{4+\delta} < \infty \text{ and } E||x_j||^{4+\delta} < \infty \text{ for some } \delta > 0.$$

The moment conditions in Assumption 1 are needed for various reasons, cf. Bierens (1987). In particular, since Assumption 1 implies $E|y_j| < \infty$ the conditional expectation of y_j relative to x_j is well-defined as a (Borel measurable) real function g of x_j:

$$E(y_j|x_j) = g(x_j),$$

Cf. Chung (1974, Theorems 9.1.1 and 9.1.2). Denoting

$$u_j = y_j - g(x_j),$$

we then get the regression model

$$y_j = g(x_j) + u_j,$$

where by construction the error term u_j satisfies the usual condition that its conditional expectation relative to the vector of regressors equals zero with probability 1 (w.p.1), i.e.

$$E(u_j|x_j) = 0 \text{ w.p.1.}$$

The model is therefore purely tautological in that its set up is merely a matter of definition. Since this definition of u_j does not imply independence of u_j and x_j, the errors u_j are in general conditionally heteroscedastic, i.e.

$$P[E(u_j^2|x_j) = Eu_j^2] < 1.$$

We assume no explicit functional form for g(.), hence the model does not contain parameters in the usual sense. In fact the function g(.) itself is the unknown "parameter" to be estimated from the data.

The next assumption describes the mixed continuous-discrete character

of the vector of regressors.

Assumption 2. Let $x_j = (x_j^{(1)}, x_j^{(2)})' \in X_1 \times X_2$, where X_1 is a k_1-dimensional real space and X_2 is a countable subset of a k_2-dimensional real space. (Thus $k_1+k_2=k$). The set X_2 is such that-

(I) $x^{(2)} \in X_2$ implies $p(x^{(2)})=P(x_j^{(2)}=x^{(2)}) > 0$;

(II) $\sum p(x^{(2)}) = 1$, where the summation is over all $x^{(2)} \in X_2$;

(III) there exists a $\mu > 0$ such that $z_1 \in X_2$, $z_2 \in X_2$, $z_1 \neq z_2$ implies $||z_1 - z_2|| > \mu$.

(IV) For each $x^{(2)} \in X_2$ the distribution of $x_j^{(1)}$ relative to the event $x_j^{(2)}=x^{(2)}$ is absolutely continuous with conditional density function $h(x^{(1)}|x^{(2)})$.

The conditions (I), (II) and (IV) speak for themselves, but condition (III) may need some explanation. It is slightly stronger than the corresponding condition in Bierens (1987, Assumption 3.2.1). Nevertheless it is satisfied for the empirical application under review. It says that distinct discrete regressors have a non-zero minimum distance, so that limit points in X_2 are excluded.

Although we do not assume an explicit functional form for $g(x^{(1)},x^{(2)})$ and $h(x^{(1)}|x^{(2)})$ we do need some regularity conditions. These regularity conditions employ the following definition.

Definition 1. Let $D_{k,m}$ be the class of all continuous real functions f on R^k such that the derivatives

$$(\partial/\partial z_1)^{i_1}(\partial/\partial z_2)^{i_2}\ldots.(\partial/\partial z_k)^{i_k}\ f(z_1,\ldots,z_k),\ i_j \geq 0\ ,j=1,..,k,$$

are continuous and uniformly bounded for $0 \leq i_1+i_2+..+i_k \leq m$.

Denote for $\varepsilon > 0$ and $x = (x^{(1)\prime}, x^{(2)\prime})' \in X_1 \times X_2$,

$$\sigma^{\varepsilon}(x) = \sigma^{\varepsilon}(x^{(1)}, x^{(2)}) = E[|u_j|^{\varepsilon} | (x_j^{(1)} x_j^{(2)} = (x^{(1)}, x^{(2)})],$$

Assumption 3. For each fixed $x^{(2)} \in X_2$ we have:

(I) The functions $h(x^{(1)}|x^{(2)})$ and $g(x^{(1)}, x^{(2)})h(x^{(1)}|x^{(2)})$ belong to the class $D_{k_1, m}$ for some $m \geq 2$.

(II) The function $\sigma^4(x^{(1)}, x^{(2)})h(x^{(1)}|x^{(2)})$ is uniformly bounded on X_1.

(III) The function $g(x^{(1)}, x^{(2)})^2 h(x^{(1)}|x^{(2)})$ has continuous and bounded second derivatives with respect to the components of $x^{(1)}$.

(IV) The matrix $V = Ex_j x_j' - (Ex_j)(Ex_j)'$ is non-singular.

These are all the assumptions we need.

A kernel estimator of $g(x)$ is now a random function of the form

$$\hat{g}(x) = \frac{\sum_{j=1}^{n} y_j K((x - x_j)/\lambda_n)}{\sum_{j=1}^{n} K((x - x_j)/\lambda_n)},$$

where K is a real function on $X_1 \times X_2$ called the *kernel*, and (λ_n) is a sequence of *window width* parameters converging to zero. For certain specifications of K and λ_n the kernel regression estimator is consistent and pointwise asymptotically normally distributed. For example, if K is specified as the density of the $k_1 + k_2$-variate normal distribution with zero mean vector and nonsingular variance matrix and λ_n is such that

$$\lambda_n \to 0, \quad \lambda_n^{k_1} \sqrt{n} \to \infty$$

then under Assumptions 1-3,

$$\sqrt{(n\lambda_n^{k_1})}[\hat{g}(x) - g(x)] \to N[b(x), (\sigma^2(x)/h(x))\int_{X_i} K(z_1,0)^2 dz_1] \text{ in distr.,}$$

[cf. Bierens (1987)], where b(x) is the asymptotic bias and

$$h(x) = h(x^{(1)}|x^{(2)})p(x^{(2)}).$$

Moreover, in this case the rate of convergence is maximal for

$$\lambda_n = c.n^{-1/(4+k_1)},$$

where $c > 0$ is a constant. This rate of convergence can be further increased by choosing a more general class of kernels.

Bierens (1987) proposes a modification of the kernel regression method in order to get rid of the asymptotic bias, to make the kernel regression approach invariant for linear transformations of the data and to get a rate of convergence in distribution arbitrarily close to $\sqrt{n}$. First, Bierens advocate the following data dependent kernel. Let for m=2,4,6,..

$$\hat{K}_m(x) = \sum_{i=1}^{m/2}\theta_i \exp(-\tfrac{1}{2}x'\hat{V}^{-1}x/\sigma_i^2)/\{(\sqrt{(2\pi)})^{k_1}|\sigma_i|^{k_1}\sqrt{\det((\hat{V}^{(1)})^{-1})}\} \quad (4.1)$$

where

$$\hat{V} = (1/n)\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})' \text{ with } \bar{x} = (1/n)\sum_{j=1}^{n}x_j,$$

$\hat{V}^{(1)}$ is the upper-left $k_1 \times k_1$ submatrix of $\hat{V}^{-1}$

and the θ_i and σ_i are such that

$$\sum_{i=1}^{m/2}\theta_i\sigma_i^{2\ell} = 1 \text{ if } \ell = 0,$$
$$= 0 \text{ for } \ell = 1,2,..,(m/2)-1.$$

In the empirical application under review we have chosen:

$$\sigma_i = \sqrt{i},\ (i=1,2,..)\ ;\ m = 8.$$

Next, let for $\lambda > 0$ and $c > 0$,

$$\begin{aligned}
&\hat{g}^*_m(x|\lambda) = \textstyle\sum_{j=1}^n y_j \hat{K}_m((x-x_j)/\lambda) \ /\ \sum_{j=1}^n \hat{K}_m((x-x_j)/\lambda),\\
&\hat{g}_{1m}(x|c) = \hat{g}^*_m(x|c.n^{-1/(2m+k_1)}),\\
&\hat{g}_{2m}(x|c) = \hat{g}^*_m(x|c.n^{-0.5/(2m+k_1)}).
\end{aligned} \tag{4.2}$$

Then the modified kernel regression function estimator of $g(x)$ as proposed by Bierens (1987) takes the form

$$\hat{g}_m(x|c) = \{\hat{g}_{1m}(x|c) - n^{-0.5m/(2m+k_1)}.\hat{g}_{2m}(x|c)\} \ /\ \{1-n^{-0.5m/(2m+k_1)}\}. \tag{4.3}$$

Denoting

$$K_m(x) = \operatorname{plim}_{n\to\infty} \hat{K}_m(x)$$

we now have:

Theorem 1. Let Assumptions 1-3 hold.

(I) For each $x \in X_1 \times X_2$ with $h(x) > 0$ and each constant $c > 0$,

$$n^{m/(2m+k_1)}[\hat{g}_m(x|c) - g(x)] \to N[0,\sigma^2_{g,m}(x|c)] \text{ in distr.}$$

where

$$\sigma^2_{g,m}(x|c) = c^{-k_1}\{\sigma^2(x)/h(x)\}\int_{X_1} K_m(z_1,0)^2 dz_1 .$$

(II) Let $x^*_1,\ldots,x^*_M$ be distinct points in $X_1 \times X_2$ for which $h(x^*_i) > 0$. Then

$$n^{m/(2m+k_1)}\begin{bmatrix} \hat{g}_m(x^*_1|c) - g(x^*_1) \\ \vdots \\ \hat{g}_m(x^*_M|c) - g(x^*_M) \end{bmatrix} \to N_M[0,\Sigma_m(c)] \text{ in distr.}$$

where $\Sigma_m(c)$ is an $M \times M$ diagonal matrix with diagonal elements

$\sigma^2_{g,m}(x_1^*|c), \ldots, \sigma^2_{g,m}(x_M^*|c)$.

This implies that the components of the random M-vector involved are asymptotically independent.

Proof: Bierens (1987).

A consistent estimator of the asymptotic variance $\sigma^2_{g,m}(x|c)$ is

$$\hat{\sigma}^2_{g,m}(x|c) = \frac{c^{-k_1}(1/n)\sum_{j=1}^n (y_j - \hat{g}_m(x|c))^2 \hat{K}_m[(x-x_j)/\lambda_n(c)]^2/\lambda_n(c)^{k_1}}{((1/n)\sum_{j=1}^n \hat{K}_m[(x-x_j)/\lambda_n(c)]/\lambda_n(c)^{k_1})^2} \qquad (4.4)$$

where

$$\lambda_n(c) = c.n^{-1/(2m+k_1)}.$$

Thus:

Theorem 2. Under Assumptions 1-3,

$$\text{plim}_{n\to\infty} \hat{\sigma}^2_{g,m}(x|c) = \sigma^2_{g,m}(x|c).$$

Proof: Similarly to Bierens (1987, form. (5.3.3)).

Combining Theorems 1 and 2 we now see that the asymptotic 95% confidence interval for g(x) is given by

$$[\hat{g}(x) - 1.96\hat{\sigma}_{g,m}(x|c)/n^{.5m/(2m+k_1)},\ \hat{g}(x) + 1.96\hat{\sigma}_{g,m}(x|c)/n^{.5m/(2m+k_1)}]$$

4.2 *Sample selection*

We recall that the data set on which the nonparametric regression results were based is a further subsample of size 1130 from a subsample of size 2000. The latter subsample is a random sample, but the former is obtained by deleting the households with only one parent or adult and the households with one or two persons in the age group 65 or over, and is therefore not a random sample. In this subsection we show now how to account for this sample selection.

Let the original *random* sample be

$$((\tilde{y}_1,\tilde{x}_1,\tilde{z}_1),\ldots,(\tilde{y}_N,\tilde{x}_N,\tilde{z}_N),$$

where $\tilde{y}_j$ is the dependent variable, $\tilde{x}_j$ is a k-vector of regressors and $\tilde{z}_j$ is a dummy variable taking the values 0 or 1. In the empirical application under review we have $N = 2000$, $\tilde{y}_j$ is one of the two expenditure categories, $\tilde{x}_j = (\tilde{x}_{1j},\tilde{x}_{2j},\tilde{x}_{3j})'$ with

$\tilde{x}_{1j}$ = net income,
$\tilde{x}_{2j}$ = number of children in the age group 0-15,
$\tilde{x}_{3j}$ = number of children in the age group 16 or over

and

$\tilde{z}_j = 0$ for households with only one adult (parent) or with one or two persons in the age group 65 or over,
$\tilde{z}_j = 1$ for other households.

We now assume:

Assumption 4. Assumptions 1-3 hold for this random sample (reading $y_j=\tilde{y}_j$, $x_j=(\tilde{x}_j,\tilde{z}_j)'$, k=k+1, n=N).

We are interested in estimating the conditional expectation

$$g(x) = E(\tilde{y}_j \,|\, \tilde{x}_j = x,\ \tilde{z}_j = 1).$$

Now let $((y_1,x_1),\ldots,(y_n,x_n))$ be a further subsample of size n corresponding to the data points $(\tilde{y}_j,\tilde{x}_j)$ for which $\tilde{z}_j=1$. Calculate the modified kernel regression estimator $\hat{g}_m(x|c)$ and the variance estimator $\hat{\sigma}^2_{g,m}(x|c)$ as if this further subsample would obey assumptions 1-3. Then the results in Theorems 1-2 go through, except that the rate of convergence in distribution now depends on N rather than on n. Thus:

Theorem 3. Let Assumption 4 hold. Let $h(x^{(1)}|x^{(2)})$ be the conditional density of $\tilde{x}_j^{(1)} \in X_1$ relative to the event

$$(\tilde{x}_j^{(2)},\tilde{z}_j) = (x^{(2)},1) \in X_2\times\{0,1\}.$$

Moreover, let

$$p(x^{(2)}) = P\{\tilde{x}_j^{(2)}=x^{(2)},\tilde{z}_j=1\},\ h(x)=h(x^{(1)}|x^{(2)})p(x^{(2)}).$$

I) For every x with $h(x) > 0$ and each constant $c > 0$,

$$N^{m/(2m+k_1)}[\hat{g}_m(x|c) - g(x)] \to N\{0,\sigma^2_{g,m}(x|c)\} \text{ in distr.},$$

where

$$\sigma^2_{g,m}(x|c) = \operatorname{plim}_{N\to\infty}\hat{\sigma}^2_{g,m}(x|c).$$

II) Let $x_1^*,\ldots,x_M^*$ be distict points for which $h(x_i^*) > 0$. Then

$$N^{m/(2m+k_1)}\begin{bmatrix}\hat{g}_m(x_1^*|c) - g(x_1^*)\\ \vdots \\ \hat{g}_m(x_M^*|c) - g(x_M^*)\end{bmatrix} \to N_M[0,\Sigma_M(c)] \text{ in distr.}$$

where $\Sigma_M(c)$ is the diagonal matrix in Theorem 1(II).

Proof: Let $\hat{K}_m(x)$ be the kernel calculated on the basis of the subsample of size n. Cf. (4.1). Define

$$\hat{K}_m^*(x,z) = \hat{K}_m(x).I(z = 0),$$

where I(.) is the indicator function, and let

$$\hat{g}_m^{**}(x,z|\lambda) = \frac{\sum_{j=1}^n \tilde{y}_j \hat{K}_m^*[(x-\tilde{x}_j)/\lambda,(z-\tilde{z}_j)/\lambda]}{\sum_{j=1}^n \hat{K}_m^*[(x-\tilde{x}_j)/\lambda,(z-\tilde{z}_j)/\lambda]}. \qquad \text{(cf.(4.2))}$$

Moreover, define

$$\hat{g}_m^*(x,z|c) \text{ and } \hat{\sigma}^{2*}_{g,m}(x,z|c)$$

similarly to (4.3) and (4.4), respectively. Then it is not hard to show along the lines in Bierens (1987) that the results in Theorems 1 and 3 go through. The theorem now follows from the fact that

$$\hat{g}_m(x|c) = \hat{g}_m^*(x,1|c),\ \hat{\sigma}^2_{g,m}(x|c) = \hat{\sigma}^{*\,2}_{g,m}(x,1|c).$$

4.3 *Choosing the constant c*

In Bierens (1987) it is advocated to choose the constant c of the window width by cross-validation. In the cross-validation approach each y_j in the sample of size n is predicted by the kernel regression estimator based on the remaining n-1 observations. Thus let $\hat{g}_m^{(\ell)}(x|c)$ be the kernel regression estimator based in the sample with the ℓ-th observation left out. Then c is determined by minimizing

$$\sum_{\ell=1}^n [y_\ell - \hat{g}_m^{(\ell)}(x|c)]^2$$

over an interval $[c_1,c_2]$, $0 < c_1 < c_2 < \infty$. A drawback of this approach is that the resulting estimated constant $\hat{c}$, say, depends on the same sample as the kernel regression function estimator. Consequently, $\hat{g}_m(x|c)$ with c

fixed is not independent of $\hat{c}$ and hence the asymptotic normality results for $\hat{g}_m(x|c)$ need not hold for $\hat{g}_m(x|\hat{c})$. Therefore we have used the smaller random subsample of size 859 for estimating c by cross-validation. Then $\hat{c}$ is independent of the kernel regression estimator $\hat{g}_m(x|c)$ based in the random subsample of size 2000 and therefore all the asymptotic normality results carry over to $\hat{g}_m(x|\hat{c})$.

The resulting cross-validated $\hat{c}$, however, appeared to be too large, by which the the kernel regression estimator became almost constant. Therefore we have conducted various experiments with alternative values of c, still confining the analysis to the smaller subsample. It appeared that the best choice for c was $\hat{c} = 2$; best in the sense that for this c the kernel regression estimate was sufficiently smooth without being flat. Using $c = 2$ the nonparametric regression analysis has been further conducted on the basis of the larger subsample of size 2000.

REFERENCES:

Amsterdam scale (1917). Arbeidersbudgets gedurende de crisis. 's-Gravenhage: Departement van Landbouw, Nijverheid en Handel.

Atwater,W.O. (1895). American Food Materials. Bulletin 28, U.S. Departement of Agriculture.

Atwater,W.O. and D.C.Wood (1898). Dietary Studies in New York City in 1895 and 1896. Bulletin 46, U.S. Department of Agriculture.

Barten,A.P. (1964). Family Composition, Prices and Expenditure Patterns. in:Econometric Analysis for National Economic Planning, edited by P.E. Hart, F. Mills, and J.K. Whitaker, London: Butterworths.

Beveridge,W. (1942). Social Insurance and Allied Services. Cmmd 6406, HMSO.

Bierens,H.J. (1987). Kernel Estimators of Regression Functions. in: Advances in Econometrics 1985, 99-144, edited by T.F. Bewley, New York: Cambridge University Press.

Blundell,R.W. (1980). Estimating Continuous Consumer Equivalence Scales in an Expenditure Model with Labour Supply. European Economic Rewiew, 14, 145-157.

Blundell,R.W. and I.Walker (1982). Modelling the Joint Determination of Household Labour Supplies and Commodity Demands. Economic Journal, 92, 351-364.

Blundell,R.W. and I.Walker (1984). A Household Production Specification of Demographic Variables in Demand Analysis. Economic Journal, 94, 59-68.

Bojer,H. (1977). The Effect on Consumption of Household Size and Composition. European Economic Review, 9, 169-193.

Brown,A. and A.S.Deaton (1972). Surveys in Applied Economics: Models of Consumer Behaviour. Economic Journal, 82, 1145-1236.

Chung,K.L. (1974). A Course in Probability Theory. New York: Academic Press.

Cramer,J.S. (1969). Empirical Econometrics. Amsterdam: North-Holland.

Deaton,A.S. (1981). Three Essays on a Sri Lankan Household Survey. Living Standards Measurement Study Working Paper 11. Washington: World Bank.

Deaton,A.S. and J.Muellbauer (1980). Economics and Consumer Behavior. New York: Cambridge University Press.

Deaton,A.S. and J.Muellbauer (1986). On Measuring Child Costs: With Applications to Poor Countries. Journal of Political Economy, 94, 720-744.

Denmark scale (1900). Danske Arbejderfamiliers Forbrug. Kobenhavn: Statistiske Meddelelser.

Diewert,W.E. (1971). Application of the Shepherd Duality Theorem: Generalized Leontief Production Function. Journal of Political Economy, 79, 481-507.

Dublin,L.I. and A.J.Lotka (1946). The Money Value of a Man. New York: Ronald Press.

Engel,E. (1883). Der Werth des Menschen. I. Teil. Der Kostenwerth des Menschen, Berlin. Reprinted as:

Engel,E. (1995). Die Lebenskosten Belgischer Arbeiterfamilien fruher und jetzt. International Statistical Institute Bulletin, 9.

Forsyth,F.G. (1960). The Relationship between Family Size and Family Expenditure. Journal of the Royal Statistical Society, Series A 123, 367-397.

Gorman,W.M. (1976). Tricks with Utility Functions. in: Essays in Economic Analysis, edited by M.Artis and A.R.Nobay. Cambridge: Cambridge University Press.

Henderson,A.M. (1949). The Costs of Children, Part I. Population Studies 3, 130-150.

Kapteyn,A. and B.M.S. van Praag (1976). A New Approach to the Construction of Family Equivalence Scales. European Economic Review, 7, 313-335.

König,J. (1907). Chemie der menschlichen Nahrungs- und Genussmittel. In J.J.R.Moquette, Onderzoekingen over volksvoeding in de Gemeente Utrecht. Utrecht.

Kuhna (1894). Die Ernahrungsverhaltnisse der industriellen Arbeiterbevolkerung in Oberschlesien. Leipzig.

Leser,C.E.V. (1963). Forms of Engel Functions. Econometrica, 31, 694-703.

Muellbauer,J.(1975). The Cost of Living and Taste and Quality Change. Journal of Economic Theory, 10, 269-283.

Muellbauer,J. (1977). Testing the Barten Model of Household Composition Effects and the Costs of Children. Economic Journal, 87, 460-487.

Muellbauer,J. (1980). The Estimation of the Prais-Houthakker Model of Equivalence Scales. Econometrica, 48, 153-176.

Nasse,R. (1891). Ueber die Haushaltung der Bergarbeiter im Saarbruckenschen und in Grossbrittannien. Jahrbucher fur Nationalokonomie und Statistik.

Nicholson,J.L. (1949). Variations in Working Class Family Expenditure. Journal of Royal Statistical Society, Series A 112, 359-411.

Oishansky,M. (1965). Counting the Poor: Another look at the Poverty Profile. Social Security Bulletin, 28.

Oishansky,M. (1968). The Shape of Poverty in 1966. Social Security Bulletin, 31.

Pollak,R.A. and T.J.Wales (1978). Estimation of Complete Demand Systems from Household Budget Data: The Linear and Quadratic Expenditure Systems. American Economic Review, 68, 348-359.

Pollak,R.A. and T.J.Wales (1981). Demographic Variables in Demand Analysis. Econometrica, 49, 1533-1551.

Prais,S.J. (1953). The Estimation of Equivalent Adult Scales from Family Budgets. Economic Journal, 63, 791-810.

Prais,S.J. and H.S.Houthakker (1955). The Analysis of Family Budgets. Cambridge: Cambridge University Press.

Ray,R. (1983). Measuring the Costs of Children: An Alternative Approach. Journal of Public Economics, 22, 89-102.

Ray,R. (1985). A Nested Test of the Barten Model of Equivalence Scales. Economics Letters, 17, 411-412.

Rothbarth,E. (1943). Note on a Method of Determining Equivalent Income for Families of Different Composition. App. 4 in: C.Madge. War-Time Pattern of Saving and Spendings, Cambridge: Cambridge University Press.

Rowntree,B.S. (1901). Poverty: A Study of Town Life. McMillan.

Rowntree,B.S. (1941). Poverty and Progress. Longmans.

Swedish scale (1908). Statistisk Undersokning angande Lefnadskostnaderne i Stockholm aren 1907-1908. X. Specialunderokninbar No. 1, Stockholm: Stockholm Bureau of Statistics.

Stone,J.R.N. (1954). Linear Expenditure Systems and Demand Analysis: an Application to the Pattern of British Demand. Economic Journal, 64, 511-527.

Sydenstricker,E. and W.I.King (1921). The Measurement of Relative Economic Status of Families. Quarterly Publication, American Statistical Association, 17, 842.

Working,H. (1943). Statistical Laws of Family Expenditure. Journal of the American Statistical Association, 38, 43-56.