University of Wisconsin-Madison

December 1995 No.389

Issues in Data Management of Expenditure Surveys: an Example from the Colombian 1984-85 Urban Survey

By

C. Federico Perali and Thomas L. Cox

AGRICULTURAL ECONOMICS

STAFF PAPER SERIES

December, 1995 **Staff Paper No. 389**

Issues in Data Management of Expenditure Surveys:

An Example from the Colombian 1984-85 Urban Survey

by

C. Federico Perali and Thomas L. Cox[1]

Abstract: This study examines some of the issues that are often encountered when analysing the socio-economic information of expenditure surveys using the Colombian 1985 urban expenditure survey as an example. This study discusses the problem of estimating prices when only information on expenditures and demographic characteristics is available, the options available when aggregating commodities and defining demographic or labour information, and the advantages and drawbacks in choosing total expenditure as a proxy for income or as the metric for poverty measurement. The study also applies non-parametric density estimation as a form of exploratory data analysis and as a means to learn from the data about the most proper parametric specification.

December 1995

---

[1] C. Federico Perali, Istituto di Economia e Politica Agraria, Universita` degli Studi di Verona and Thomas L. Cox, Department of Agricultural Economics, University of Wisconsin, Madison. The authors wish to thank, without implicating, Jean-Paul Chavas and Maria Luisa Ferreira for their helpful comments and discussion.

I. Introduction

This document deals with some issues crucial in managing household expenditure surveys. Researchers managing expenditure surveys are confronted with decisions about how to estimate and aggregate prices, the most appropriate and economically relevant commodity grouping, the statistical representation of demographic information, and how to choose between total expenditure and income in order to understand consumption behaviour as well as improving the precision of the measurement of poverty. The study also introduces some popular non-parametric techniques to estimate densities and Engel relationships as a tool for controlling the accuracy of the data management and as a complement to classical parametric methods. These issues will be treated in order.

The 1985 Colombian urban survey is used as a reference sample from which examples are drawn. The simple descriptive statistical analysis proposed describes information generally available in the majority of Latin America Expenditure Surveys. These procedures can be used for making the relevant economic information uniform across surveys and hence, to facilitate the access to the cross-country survey information by econometricians.

II. Relevant Information About Expenditure Surveys

This section provides a summary, though not exhaustive, of the general information usually provided in the documentation available from the statistical institutions responsible for the collection of the survey and is of importance to the applied researcher. This example is related to the 1985 Expenditure Survey of Colombia.

The *Encuesta Nacional de Ingresos y Gastos de Colombia, 1984-85*" (Colombian Income and Expenditure Survey - CIES) was conducted during the period March 1984 and February 1985 by the "Departamento Administrativo Nacional de Estadistica - DANE." The survey was undertaken to update the composition of the basic consumption of the household basket, to increase the level of detail and

extend the geographic base from seven to thirteen cities. It constitutes the basis for the implementation of the new Colombian consumer purchasing power index known as the IPC-60 which is still adopted.

The survey covered urban areas only. Households have been sampled in the equal proportion during the four seasons. The cities included in the survey along with the number of households surveyed in each city are reported in Table 1.

Both the daily and the less frequent expenditures were collected during weekly interviews using the "booklet method" according to which the head of the sampled households were asked to fill the questionnaire under the supervision of the interviewer. The questionnaires measure quantities in the units usually adopted for each item such as litres for beverages, bags of 1 litre for milk, pounds for salt and meat, dozens for oranges with the associated cash expenditures. The tapes, however, report all quantities in grams. The questionnaire for less frequent expenditures provide also the total amount paid with cash or credit. The survey refers to goods purchased, not to goods consumed.

The daily expenditures are mainly composed of prepared and non-prepared food, food eaten away from home, alcoholic and non-alcoholic beverages, cigarettes, fuel for domestic use and gasoline for the family car, newspapers, lotteries, urban transportation, telephone and others. Food not acquired through market transactions such as food coming from family activities like grocery shops, bakeries or farming is also accounted using market prices to estimate their total value.

Non-food expenditures refer to the non-frequent expenses with monthly, trimestral and annual periodicity incurred in the month, trimester or year preceding the week of interview. For examples, vehicles, motorcycles, bicycles, taxes on the vehicle have annual periodicity, while clothing expenses have been recorded relative to the past three months. This is a distinctive characteristic of the Colombian expenditure survey. For example, Pudney (1990) reports that the 1983 UK Family Expenditure Survey (FES) is based on a two-week observation period with no recall. In the discussion to Pudney's paper, John Muellbauer (in Myles, Editor, 1990: 307) remarks on this aspect by pointing out that: " ... in many surveys

in other countries, clothes expenditure is taken over a three month interval and the refrigerators over a one-year recall. The FES could do that." This observation is important, because it allows interpreting zero expenditures as the expression of genuine non-consumption rather than as the outcome of infrequent purchases.

The length of the recall period for reporting transactions is critical in the design of household expenditure surveys. The longer the recall period, the higher is the probability of recording recall errors. However, the longer the recall period, the better is the estimation of the expected long-term transactions and the regularity of purchases and consumptions. This trade-off is particularly important to survey designers in less-developed countries where it is often difficult to implement the diary method due to high illiteracy rates in the local population. Scott and Amenuvegbe (1990) report that in most African surveys, food expenditures have been recorded by interviewers visiting either every day or every second day. In Asia, recall periods of one week or one month have been more common. The authors argue that the choice of the recall period is the most urgent design issue facing third world survey workers today.

III. The Estimation of Prices

This section first describes the methodology to derive unit values and discusses some related empirical issues as emerged from the application to the case of the Colombian data set, then introduces the theory and the method underlying the derivation of price information using the demographic characteristics of the household. A brief discussion of price aggregation issues and the treatment of seasonal and regional price trends concludes the section.

Some recent expenditure surveys (e.g., the Mexican 1984 and 1989) ask directly for the prices paid by each family during the interview period and then expenditures on each commodity are deduced from the direct knowledge of the quantities purchased. For such surveys the estimation of prices are not an issue. In general, however, prices are deduced from the knowledge of expenditures and quantities

expressed in a common unit. These implicit prices are more properly referred to as unit values. In other cases, only expenditures are recorded. In these cases household-specific prices may be recovered using household-specific demographic information as suggested by Lewbel (1989).

Define the unit value u_i of a commodity i as the implicit price paid per physical unit. The unit values for the j=1,.., m groups of non-durables goods have been computed as a weighted average:

$$U_j = \sum_{i=1}^{n} w_i \, u_i \,, \quad for \quad i=1,..,n \quad and \quad j=1,..,m, \qquad \textbf{(1)}$$

where the weights $w_i = (y_i/y_j)$ are the shares of each good i over the total expenditure y_j for group **j**. Unit values refer in average to August 1984. Note that for durable goods and for food away from home, all quantities are assumed to be equal to one, so that $u_i=y_i$. As a consequence, the unit value for the groups of durable goods becomes:

$$U_j = \sum_{i=1}^{n} w_i \, y_i \,, \quad for \quad i=1,..,n \quad and \quad j=1,..,m \,. \qquad \textbf{(2)}$$

This assumption is not offensive because most durable goods are bought in single units. In relation to non-frequent purchases such as clothing that can be bought in multiple units, we regard the monthly expenditure on clothing as a single unit of an undifferentiated good. In this sense, the number of bus, taxi, train, etc. trips per month corresponds to one unit of monthly transportation. The number of litres of gasoline, kitchen and heating fuel, kilowatts of electricity, etc. corresponds to one unit of the monthly consumption of the commodity "energy."

The quantity of food items may not be available and hence, unit values cannot be computed. In the Colombian case, the food items reported without quantities were *masa*, bread, *arepa*, *bollos* and *envueltos*, salt, corn flakes, other cereales, *manos, patas o cabeza de res*, eggs, aromatic herbs, *aji*, cakes and sweeties, fried potatoes, the, ice-creams ready-to-eat, other aromatic herbs, and ice in blocks. Given

their importance in the Colombian diet, only eggs and bread received special treatment. The price for one egg was taken from the market prices in each city published by DANE for August 1984 in the *Boletín de Estadística* series. The price for eggs was maintained on a per unit basis because the expenditures were reported based on the same unit. The price for bread was not available for each city in the published statistics. It was derived from the May 1990 statistics and converted to 1984 prices using the Colombian CPI for food with base 1978.[2] This correction is important in order to ensure at least comparability across levels of relative aggregate prices. The household variability of the commodity's unit value is inevitably lost, and the variability of the aggregate unit value of the commodity group to which such commodities belong is underestimated. Surveys reporting only expenditure information are traditionally used in Engel curve analysis leading to the estimation of income elasticities. Price elasticities can be recovered resorting to very restrictive assumptions (Frisch 1959) generally rejected by statistical hypotheses tests in empirical applications. Furthermore, these surveys have less usefulness in modern welfare and policy analysis. If price information were not recoverable at all, then expenditure surveys that do not record quantity information are less useful for studies comparing economic results across countries and/or across time with the requirement of applying the same econometric technique to uniform data sets.

The following methodology proposed by Lewbel (1989) uses generalized "within-group" equivalence scales, defined here as the ratio of the group sub-utility function to the corresponding sub-utility function of a reference household, estimated without price variation in place of "between-group" price variation. The method relies on the assumption that the original function is homothetically separable and "within-group" sub-utility functions are Cobb-Douglas.

Consider a separable utility function $U(u_1(q_1,d),..,u_n(q_n,d))$ where $U(u_1,..,u_n)$ is the "between-

---

[2] The national price for one egg in August 1984 was 9.03 Colombian pesos. The computed national price for one kilogram of bread in the same month, excluding the outlier price in Villavicencio of 192.09 Colombian pesos per kilogram, was 151.45 Colombian pesos. This figure compares well with the published price for one kilogram of bread of 138.82 Colombian pesos (DANE, *Boletín de Estadistica*, 429/December 1988:153).

group" utility function and $u_i(q_i,d)$ is the "within-group" sub-utility function. The index i=1,..,n denotes the aggregate commodity groups while n_i is the total number of goods q comprising group i. Demographic characteristics d affect U(.) only indirectly through the direct effects on the within-group sub-utility function. By the definition given above, define the group equivalence scale $M_i(q,d)$ as:

$$M_i(q,d) = \frac{u_i(q,d)}{u_i(q,d^r)} \qquad \textbf{(3)}$$

where d^r describes the demographic profile of a reference household. Define a quantity index for group i as $Q_i(u_i,d^r)$ and rewrite the between-group utility function as:

$$U(u_i,..,u_n) = U\left(\frac{Q_1}{M_1},..,\frac{Q_n}{M_n}\right) \qquad \textbf{(4)}$$

which is formally analogous to Barten's (1964) technique to introduce demographic factors in the utility function. Define further the price index for group i as $P_i = Y_i^r/Q_i$ where Y_i^r is expenditure on group I by the reference household. Barten's utility structure implies the following share demands for each household:

$$W_i = H_i(P_1M_1,..,P_nM_n,Y) \qquad \textbf{(5)}$$

taking the form of $W_i^r = H_i(P_1,..,P_n,Y^r)$ for the reference household with scales $M_i = 1$ (Pollak and Wales 1981). The further assumption of homothetic separability admits two-stage budgeting and implies the existence of indirect sub-utility functions V_i such that $P_i = V_i(p_i,d^r)$. By analogy with the definition of group equivalence scales in utility space, it follows that:

$$M_i = \frac{V_i(p_i,d)}{V_i(p_i,d^r)} \qquad \textbf{(6)}$$

and $V_i = M_i P_i$. Therefore, when demands are homothetically separable each group scale depends only on relative prices within group I and on d as required (and assumed).

Maximization of $u_i(q_i,d)$ subject to the expenditure in group I $p_i q_i = x_i$ gives the budget share for an individual good $w_{ij} = h_{ij}(p_i,d,x_i)$. For homothetically separable demands, then the budget shares do not depend on expenditure $w_{ij} = h_{ij}(p_i,d)$ and integrate back in a simple fashion to $V_i = M_i P_i$. This information can be used at the between-group level in place of price data to estimate $W_i = H_i(V_1,..,V_n,X)$.

Under the assumption that the sub-group utility functions are Cobb-Douglas with parameters specified as "shifting" functions of demographic variables alone, i.e.,

$$F_i(q_i,d) = k_i \prod_{j-1}^{n_i} q_{ij}^{m_{ij}(d)}, \qquad \textbf{(7)}$$

then the shares correspond to the demographic functions:

$$w_{ij} = h_{ij}(p_i,d) = m_{ij}(d) \qquad with \qquad \sum_{j-1}^{n} w_{ij}(d) = \sum_{j-1}^{n} m_{ij}(d) = 1. \qquad \textbf{(8)}$$

The implied indirect utility function is:

$$V_i(p_i,d) = M_i P_i = \frac{1}{k_i} \prod_{j-1}^{n_i} \left(\frac{p_{ij}}{m_{ij}} \right)^{m_{ij}(d)} \qquad with \qquad k_i(d) = \prod_{j-1}^{n_i} m_{ij}(d^r)^{-m_{ij}(d^r)} \qquad \textbf{(9)}$$

where $k_i(d)$ is a scaling function depending only on the choice of the reference demographic levels.

These results support a simple procedure to estimate price variation in survey data without quantity information. Jointly estimate the m_{ij} equations and the fitted shares using the stochastic specification $\hat{w}_{ij} = \hat{h}_{ij} = m_{ij}(d) + u_{ij}$ where u is a spherical error term for the within-group budget shares. Then, further assuming with no loss of information, that $p_{ij} = P_i = 1$ for all i and j, price information can be deduced from demographic information alone by using (9):

$$M_i P_i = M_i = \frac{1}{\hat{k}_i} \prod_{j-1}^{n_i} \left(\frac{1}{\hat{m}_{ij}} \right)^{\hat{m}_{ij}} = \frac{1}{\hat{k}_i} \prod_{j-1}^{n_i} m_{ij}^{-m_{ij}} \quad and \quad \hat{k}_i(d) = \prod_{j-1}^{n_i} \hat{m}_{ij}(d^r)^{-\hat{m}_{ij}(d^r)} \qquad \textbf{(10)}$$

by treating M as price data. It is important to note that the Cobb-Douglas assumption places restrictions only at the within-group level while letting the between-group demand equations free to be arbitrarily flexible. An approximation to equation (10) can be obtained by using the observed within group budget shares:

$$M_i P_i = M_i = \frac{1}{k_i} \prod_{j-1}^{n_i} w_{ij}^{-w_{ij}} \qquad \textbf{(11)}$$

and by choosing k_i according to some prior knowledge.

In general, the household-specific prices recorded in cross-sections or estimated from point-values of expenditures vary at each location and time of survey. Differences in prices experienced by different households in different areas and different time can be corrected by introducing in the estimation both seasonal and regional dummies to de-trend prices. In a study about Ghana, Glewwe and Twum-Baah propose to adjust nominal expenditures by deflating the price series prior to estimation using a monthly price index. The authors corrected for regional variation by constructing an area deflator as a regional weighted price index as $Pi_r = \sum_i w_i P_{ir}/P_{ig}$ where the weight w_i is the budget share for each good i, P_{ir} is the average price of good i in area r and P_{ig} is the average price in Ghana.

Unit values for goods that are not consumed cannot be estimated directly. Auxiliary regressions can be used to predict the unobserved unit values using a missing value technique (Dagenais 1973, Gourieroux and Montfort 1981). For example, the observed unit values can be regressed on objective factors explaining price variation such as space and time variables, income and interaction terms. Therefore, the price vector is the union of the set of estimated unit values, when a zero occurs, and the set

of actual unit values when a positive expenditure is recorded.

IV. Aggregation of Commodity and Demographic Information

Expenditure surveys generally record highly detailed commodity information. For example, the Mexican expenditure surveys record as many as 20 different type of cereals. This level of disaggregation is useful for studies interested in the estimation of second stage demand systems of small subgroups of commodities. Complete first-stage demand system are not easily estimated at the highest level of disaggregation nor they may be economically interesting. Thus, some level of aggregation is often necessary. The grouping generally assumes that preferences are weakly separable in subgroup utility functions, at a minimum. As a consequence, quantities within each group can be estimated as functions of the group expenditures and prices within the group, thus limiting the substitution possibilities to the goods belonging to the group.

The proposed aggregation of the food items in groups closely reproduces the level of aggregation that can be found in the United Nations Food and Agriculture Organization (FAO) Statistics. The definition of the food groups is as follows:

WARR = rice;
WMAI = corn;
WOCE = bread, wheat, pasta & other cereals;
WYUC = cassava;
WPAP = potatoes;
WORA = other roots;
WAZU = sugar;
WFRI = beans;
WOLE = other dried legumes;
WVEG = vegetables and other non dried legumes;
WFRU = fruit;
WCRE = red meat;
WOCA = other meat and eggs;
WPOL = poultry;
WLEC = milk and other dairy products;
WACE = fats and oils;
WAFH = food away from home.

The variable names also refer to the names of the shares.

The non-food groups have been also aggregated referring to the level of detail of the United Nations Statistics. The non-food groups are:

WVIV = HOUSING: rents, domestic fuel, public services as water, furniture, domestic appliances, etc.;
WSAL = HEALTH: medicines, therapeutical equipment, visits, hospital expenses and insurance;
WEDU = EDUCATION: discs, televisions, cinema, books, teaching services, school articles;
WVEA = ADULT CLOTHING: clothes and shoes for man and woman, cloth and shoe repair, accessories;
WVEN = CHILD CLOTHING: clothes and shoes for children and babies, repair and accessories;
WENE = ENERGY and TRANSPORTATION: car, motorcycle and bicycle purchases, transportation services;
WOTR = OTHER GOODS and SERVICES: personal articles and services, jewelry, hotels, tourism, etc.

In the Colombian survey, all quantities for food items have been converted from grams to kilos. All expenditures for durables and non-durables have been translated to a per month basis. One of the possible problems that can be encountered when aggregating commodities is that not all the goods are recorded with the same units. For example, in the Colombian survey, tobacco and packed liquid beverage, an almost insignificant fraction of all beverages, have been eliminated because quantities were reported only as units. Otherwise, a different unit of measurement other than grams or liters would have been introduced and liters or kilos would have been summed with units.

The set of demographic variables presented below is commonly present in all expenditure surveys. The variables' definition may serve as a basis to uniform the level of aggregation of the demographic characteristics across surveys.

Classes of Age for Males

AM0 = number of males with $0 \leq \text{age} < 1$
AM1 = number of males with $1 \leq \text{age} \leq 5$
AM2 = number of males with $6 \leq \text{age} \leq 10$
AM3 = number of males with $11 \leq \text{age} \leq 20$
AM4 = number of males with $21 \leq \text{age} \leq 35$

AM5 = number of males with 36 ≤ age ≤ 55
AM6 = number of males with 56 ≤ age.

Classes of Age for Females

AF0 = number of females with 0 ≤ age < 1
AF1 = number of females with 1 ≤ age ≤ 5
AF2 = number of females with 6 ≤ age ≤ 10
AF3 = number of females with 11 ≤ age ≤ 20
AF4 = number of females with 21 ≤ age ≤ 35
AF5 = number of females with 36 ≤ age ≤ 55
AF6 = number of females with 56 ≤ age.

Education Level for the Head of the Family

EA = no education
EB = ≤2 years of primary education or writing and reading capability
EC = ≥3 years of primary education and ≤2 years of secondary education
ED = ≥3 years of secundary education and ≤2 years of university
EE = ≥3 years of university.

Employment Status of the Head of the Family

TJ = 1 if head worked ≥ 6 months in previous 12 months; TJ = 0 otherwise.

Employment Status of the Wife of the Family

TS = 1 if wife worked ≥ 6 months in previous 12 months;TJ = 0 otherwise.

Occupation of the Head of the Family

EPRO = 1 if Professionals; EPRO = 0 otherwise
EMAN = 1 if Managers and Public Functionaries; EMAN = 0 otherwise
EADM = 1 if Administrators; EADM = 0 otherwise
EMER = 1 if Merchants; EMER = 0 otherwise
ETSS = 1 if Workers in the Service Sector; ETSS = 0 otherwise
ETAG = 1 if Agricultural Workers; ETAG = 0 otherwise
ETNA = 1 if Non Agricultural Workers; ETNA = 0 otherwise
ENTR = 1 if Non Workers; ENTR = 0 otherwise.

Job Position of the Head of the Family

OOBR = 1 if Worker; OOBR = 0 otherwise
OEMP = 1 if Employee; OEMP = 0 otherwise
OEMD = 1 if Domestic Employee; OEMD = 0 otherwise
OIND = 1 if Independent Worker; OIND = 0 otherwise
OPAD = 1 if Employer; OPAD = 0 otherwise
ENTR = 1 if Non Workers; ENTR = 0 otherwise.

Regions according to the Geographical Location of the Cities

R1 = 1 if COAST * Barranquilla, Cartagena; R1 = 0 otherwise
R2 = 1 if NORTH * Valledupar, Montería; R2 = 0 otherwise
R3 = 1 if BOGOTA`; R3 = 0 otherwise
R4 = 1 if ANDEAN * Bucaramanga, Medellin, Pasto,Manizales or R4 = 0

R5 = 1 if INTERANDEAN * Cali, Ibagué,Cúcuta,Pereira,Neiva or R5 = 0
R6 = 1 if ORIENTAL * Villavicencio; R6 = 0 otherwise.

Rural or Urban Location

RURAL=1 if the household lives in rural areas; RURAL=0 otherwise

Number of Persons in the Family

NPERS = Number of Persons

Numbers of Perceptores

PERCEP = Number of Perceptores

Number of Members Receiving Pension

N_PENS= Number of Members receiving Pension

Married Couple

MC_CH=1 if the household is a married couple with at least 1 child; MC_CH=0 otherwise

Single Member Household

SINGLE_H=1 if the household is composed by 1 adult; SINGLE_H=0 otherwise

Income of the Head of the Household

INGJEFE

Income of the Family

INGHOGA

Income Dummy

INGDUMY=1 if income of the head is > 0; INGDUMY=0 otherwise.

Seasons

PRIMAV=1 if Spring; PRIMAV=0 otherwise
VERANO=1 if Summer; VERANO=0 otherwise
OTONO=1 if Autumn; OTONO=0 otherwise
INVERNO=1 if Winter; INVERNO=0 otherwise.

House Ownership

ARRENDA=1 if the family rents the house; ARRENDA=0 otherwise
PROPAGA=1 if owns but still pays the house; PROPAGA=0 otherwise
PRONNPAGA=1 if the house is fully owned; PRONPAGA=0 otherwise
INVASOR=1 if the household is an invasor; INVASOR=0 otherwise.

Living Conditions

T_HOUSE= Type of house
N_ROOM= Number of rooms
WATER=1 if the house is connected with a aqueduct; WATER=0 otherwise
TOILET=1 if the house has toilet; TOILET=0 otherwise
EN_ELE=1 if the house has electric energy; EN_ELE=0 otherwise

TELEF=1 if the house has a telephone; TELEF=0 otherwise
HEAT=1 if the house has a heating system; HEAT=0 otherwise
FRIDGE=1 if the house has a fridge; FRIDGE=0 otherwise
TELEV=1 if the house has a television; TELEV=0 otherwise
VEHIC_P=1 if the household owns a car; VEHIC_P=0 otherwise
MOTO=1 if the household owns a motorcycle; MOTO=0 otherwise
O_MOTO=1 if the household owns other means of locomotion; O_MOTO=0 otherwise
ELEC_E=1 if the household owns electronic equipment; ELEC_E=0 otherwise
H_ELEC=1 if the household owns electric equipment for the house; H_ELEC=0 otherwise

Working Time

H_WORK= number of working months in the last 12 months

Working Experience

Y_WORK=years in the same occupation

The information about labour employment needs special consideration. Labour information is recorded very differently across expenditure surveys. For example, in the Colombian survey, working time is estimated by asking the number of working months in the past year. The Mexican survey asks both how many weeks was a person employed in the last month and the amount of hours worked in the past week. The Brazilian data sets asks the amount of hours worked in the actual occupation and the number of different jobs occupied each year.

The descriptive statistics of some of the variables defined above taken from the original Colombian sample are reported in Table 2, Table 3, and Table 4 both for the unconditional sample -- including both zero and positive observations -- and the conditional sample -- including only positive observations.[3] The unconditional section of the tables along with minimum and maximum values, means, and standard deviations reports also the level of truncation expressed as a percentage. Table 2 reports the shares for

[3] Previous to the computation of descriptive statistics, the observations with incomplete surveys or inconsistencies such as zero total expenditure, zero food share, or one of the shares being equal to one are eliminated from the data set. Outlier analysis is used to detect anomalous observations likely to present errors in measurement or in data input. This is not used as a means for trimming the data. The information in the tails of the distribution, particularly of income or total expenditure, are generally of high economic interest. These “extreme” observations may affect the rank of a demand system and support the introduction of quadratic terms for income or total expenditure in the Engel functions.

food items, for food (WALIM), for non-food goods, and total expenditures on all goods (GTO) and food only (GAL). Table 3 describes the price series composed by the observed values for the purchased commodities, and the estimated values for the goods not consumed. Table 4 shows the summary statistics for part of the proposed level of aggregation of the demographic information and income.

Table 5 and Figure 1 present the frequency distribution of the variable family size. This variable has been chosen as a representative distribution of the demographic information since it represents an important variable in the analysis of the cost of children. For example, the choice of a childless couple as a reference family type to be used in the computation of adult equivalence scales would represent 8.19 percent of the population.

V. Total Expenditure versus Income

The reliability of total expenditure as a proxy for income depends on the precision of the estimates of the frequency of purchase. Demand functions estimated from cross-section data are often conditional upon positive expenditure. The neglect of the behavioural information contained in the observations with zero expenditures can be a non-trivial loss; selecting on positive expenditures can significantly affect the estimate of total expenditure if a corner solution is the manifestation of a choice that needs to be explained.

Pudney enunciates at least three mechanisms explaining zero outcomes. A person may not purchase a particular good because the survey period is too short, or may not consume simply because she/he could not find the desired good, or the observed zero is the outcome of a free choice. Pudney (1990) refers to this last alternative as "economic non-consumption." The individual deliberately chooses not to consume the good given his current budget and the prices she/he faces. This behaviour represent a genuine corner solution of the individual's utility maximization problem specified in a Kuhn-Tucker framework.

The need for an exhaustive econometric treatment of the zero realizations is even more compelling in the context of poor societies, where the frequency of zero outcomes is likely to be higher relative to more affluent societies. In the case of the Colombian survey that records the purchases of durable goods referring to a reasonable period of recall, zero outcomes are more likely to be the expression of a corner solution, rather than infrequency of purchase.

Consumers facing a liquidity constraint do not buy certain goods, because they might be out of their reach. Two families facing the same prices and with the same income, but with different demographic characteristics may show very distinct consumption patterns. The relatively poorer family, in real "demographically deflated" terms, may not afford the same commodity bundle available to the richer household because it has different reservation prices or may have to trade lower quality for more quantity. Economic non-consumption is especially frequent for goods classified as luxuries. Estimates of Engel curves of rank three provide evidence for this behaviour (Blundell *et al.*, 1989). This explains why poor consumers may perceive more goods as "luxuries" relative to rich consumers.

Moreover, for policy purposes, the object of interest is demand unconditional on positive consumption. A change in price or tax policy may induce some family types to enter the consumption of certain commodities, and/or induce other family types to exit. The estimation of unconditional demands permits the computation of reservation prices which indicate how much the price would have to be lowered to induce some positive consumption.

This discussion shows that modelling the rate of consumption in developing countries is a critical issue. Household expenditure surveys generally do not record either the frequency of purchase or the frequency of consumption if purchased. It is generally assumed that the frequency of purchase during the survey period is equal to one. Suppose that the period of interview is of two weeks and that the information for durable goods is recorded with no recall. Given this setting, it is possible to observe families purchasing the same quantity of a good during the survey period, but at different rates of purchase

and consumption. This information is crucial if the objective is to test hypotheses about consumer behaviour. Such hypotheses are based on stable long-run expected rate of consumption under *ceteris paribus* conditions. For example, if the period of analysis is a month, then multiplying the quantities by two would be based on the false assumption that all families have same rate of purchase and consumption. This would lead to over-estimation of the expected long-term consumption.

In the longer run some consumers may enter consumption, some may exit. This behaviour depends on the rate of consumption which is household-specific and is also related to the type of commodity in question (e.g., its degree of storability). In general, unobserved equilibrium consumption, c, which is the choice variable of the consumer maximization problem, can be inferred using the knowledge of the expected purchasing behaviour of a collection of similar individuals assumed to be at a similar stage of the purchasing cycle.

Let P(y>0 | x,c) be the probability of one purchase realized during the survey period, where x is a set of both economic and demographic variables describing the individual. In line with Deaton and Irish (1984), Keen (1986), and Pudney (1990), unobserved consumption c is assumed to be stochastic and determined prior to purchasing behaviour. So, consumption corresponds to the expected expenditure in a hypothetical long-run, not conditional exclusively on positive expenditures:

$$c = E(y \mid x) = P(y>0 \mid x,c)\ E(y \mid y>0). \qquad \textbf{(12)}$$

This is the unconditional object of interest that relates observed expenditures to the underlying quantity demanded in the long-run. The estimation of expected behaviour corrects for erroneous inferences such as assuming that an observed zero implies that a good is not consumed, or that an observed positive expenditure is an exact estimate of underlying consumption. Some goods may not be affordable at all, and/or some may be affordable at frequency of purchase not captured by the length of the survey. Some goods may be potentially affordable, but are not purchased either because they are not available in the

market place, or the good is not found due to the search behaviour of each household. Some other goods, such as alcohol or tobacco may be conscientiously not consumed or unconsciously consumed. This suggests that the measure of consumption as expressed by the expected unconditional expenditure should have a statistical specification for each class of goods. Pudney (1990) provides a classification of goods and a good-specific statistical model for single-equation Engel curve analysis. The models are specified to ensure both the monotonicity of the Engel equations and the correct representation of the behaviour of the consumers when some goods are not consumed.

Pudney (1990) classifies goods such as tobacco and alcohol into different types characterized by conscientious abstention or a mixture of conscientious abstention and economic non-consumption. This is an important distinction under both a behavioural point of view and a metric perspective since tobacco and alcohol are clearly separable from children goods. However, in the context of poor societies, tobacco and alcohol are likely to be important in influencing the state of being of a person, but have negligible nutritional importance. For this reason, these goods are not receiving special attention in the context of the present study.

The main concern here is to incorporate in the model the behavioural content expressed by a zero realization due to economic non-consumption; that is, the type of goods not consumed due to lack of affordability are usually luxury goods. The elements of the set of goods not consumed because they are too expensive varies as income varies; i.e., the same good can be a luxury or a necessity at different income levels. For goods not consumed because of genuine economic reasons it is possible to measure the reservation price from the revealed choice of not consuming (Perali and Cox 1994). The reservation price indicates how much the market price should be reduced to induce the consumption of the good by household types of interest.

In this framework, the consumer solves the maximization problem subject to inequality constraints satisfying the Kuhn-Tucker conditions (Lee and Pitt 1986, Perali and Cox 1994). The solution generates

censored Tobit demands, such that equation (12) becomes:

$$c = E(y_i|x) = P(y_i>0|x,c)E(y_i|y_i>0) = \Phi_i(\beta'x)\ f(\beta,x_i) + \sigma\ \phi_i(\beta'x). \quad \textbf{(13)}$$

The applicability of the Tobit model depends crucially on the assumption that the zero realization is economic in nature. The robustness of this assumption can be verified by examining the distribution of the zero expenditures by income class and within income class shown in Table 6.a and Table 6.b.[4] The tables report the frequency of zeroes for some of the detailed commodities belonging to the food group such as rice (WARRZ), cassava (WYUCZ), beans (WFRIZ), milk (WLECZ), and beef (WCREZ) along with the commodities included in the econometric analysis with positive truncation. These categories are food away from home (WFAHZ), health (WSALZ), education (WEDUZ), clothing for adults (WVEAZ), clothing for children (WVENZ), and energy and transportation (WENEZ).

Tables 6.a and 6.b also report some items of the food category to provide an indication of the loss of information that aggregation over goods produces. The higher the level of aggregation across goods, the lower the economic significance of the information about the heterogeneity of consumers. For example, rice (WARRZ) is not purchased by almost 22 percent of the sample of 25644 households. Only 4.75 percent are consuming units belonging to the first income class below the indigence line. Indigent households not purchasing rice are 10 percent of all destitute families. On the other hand, the highest income level counts more than 36 percent of all the households not purchasing rice corresponding to almost 40 percent of all the households belonging to the fifth class. This may be a sign of income related behavior and preferences.

More generally, households with more income take advantage of a wider spectrum of substitution

---

[4] The first income class groups the households below or at the indigency line. The second income class includes households with income ranging between the indigency line and the poverty line. The other classes have been defined maintaining as a range the distance between the indingency and the poverty line. Both lines are as estimated by the National Administrative Department of Statistics (DANE) of Colombia in conformity with the definition provided by the Economic Commission for Latin America (CEPAL).

possibilities such that the frequency of consumption is less than one on a two weeks basis. In the case of milk (WLECZ) only 6.5 percent of the households in the sample do not buy milk. More than 75 percent of the households are concentrated below the poverty line delimited by the second income class. On the contrary, almost 66 percent of the households in the survey do not eat away from home (WFAHZ). Of these, more than 58 percent are below the poverty line. More than 70 percent of the households in the two lower income classes do not eat away from home.

Less than 10 percent of the population does not purchase health (WSALZ) or education services (WEDUZ), or clothing for adults (WVEAZ). Notice that in the case of clothing for adults (WVEAZ) it is less likely that the zeroes can be attributed to infrequency of purchase since the interviewed households were asked with a three month recall. The fact that more than 76 percent of the zero expenditures are concentrated in the low end of the income distribution says that economic non-consumption is the most plausible explanation. This evidence contrasts with the case reported by Blundell and Meghir (1987) who do not consider zero expenditures on clothing as corner solutions. It must be emphasized, that these authors use the UK Family Expenditure Survey that, besides describing an affluent society, has no recall. The results reported in Tables 6a and 6b lend support to the interpretation of the zeroes as corner solutions also in regards to the several of the other goods considered.

Clothing for children (WVENZ), on the other hand, is not bought by 9906 observations corresponding to almost 39 percent of the full sample (Table 6.a). It is reasonable to expect that the households not buying clothes for children are the ones without children. However, the disaggregation across income classes shows a different picture. In the indigent class, 12.5 percent of the households do not purchase clothing for children (WVENZ). This corresponds to almost 47 percent of the households in the lowest income class (Table 6.b), while only 14.8 percent of the households in this income class do not have children (i.e., as shown by the number of zeros in the variable counting the presence of children between 0 and 10 years of age in the household (NCH010), Table 6b). In the second income class, as

well, the percent of households not purchasing clothing for children is higher than the percent of families not having children.

If we assume stochastic independence between the variable signaling the presence for children and the purchase of clothing for children, then the probability of a zero realization is given by the product of the respective probabilities along the line of a Double-Hurdle model (Cragg 1971, Blundell and Meghir 1987). At low income levels the probability to find children in the households is close to one, so that a simple Tobit model may still be a reasonable representation.

This discussion is reversed for the other income classes. The percent of households not purchasing clothing for children is less than the percent of families without children. In the upper classes, many families do not have children (i.e., NCH010 is 40-50% for income classes 3-5 in Table 6b) . Nonetheless moany of these higher income households do buy clothes for children (i.e., 1-WENEZ is 75-95% in Table 6b). This observation also suggests a Double-Hurdle representation of behavior. However, in a demand system setting the assumption that a simple Tobit model can capture the relevant behavioral information is acceptable (i.e., the degree of censoring is only 5-25% for WENEZ in these income classes).

This discussion shows that, given the characteristics of the Colombian survey, the interpretation of zero expenditures as corner solutions, and, hence, the adoption of the Tobit model to estimate the rate of consumption, has a tenable empirical justification. This approach gives sufficient information to provide a reliable estimate of "long-run" consumption and expenditure response.

Measurement errors in estimating income arise from the imputation of values to non-objective sources of income and the mechanism adopted in the questionnaire to provide incentives to the respondent to ensure a truthful revelation of their earnings. In the case of Colombia, total income has been measured following the guidelines given by the United Nations (DANE, *Documento Metodologico de la Encuesta de Ingresos y Gastos* - 1984-1985, Bogotá, 1987) that take into account also payment in kinds, self-consumption and self-supply. These components are still very important in the Latin America context

especially in the cities of intermediate size that can be considered frontier areas with mixed urban-rural characteristics.

Personal income is determined as the sum of (a) primary income given by wages and subsidies both in cash and in kind from dependent and independent work, (b) capital income composed by imputed rents for the house if owned, rents from other properties, interests and dividends, and © current transfers such as pensions and loans from other families.

Personal available income is determined by subtracting the deductions established by law such as taxes, other contributions and social insurance from total personal income. The deductions recognized by law are accounted as expenditures. As a consequence, the study adopts exclusively the concept of total income. Total household income is computed summing the incomes of all preceptors not counting domestic employees, retired persons and excluding occasional earnings (such as inheritances, sales of properties or lottery winnings) with the estimate of the use value of the house owned, self-consumption and self-supply.

It is notorious that the reliability of the results about income measures are affected mainly by errors independent from the characteristics of the sample. During data collection, DANE's interviewers tried to explicitly correct for some of the main sources of errors. Typically, they had to cope with problems such as (a) the understatement, often involuntary because of lack of documentation, of incomes derived from independent work, and (b) the measurement of self-consumption, transactions generated under sharecropping contracts, rents in kind of capital, acquisition of food by hunting, and payments in kind in general, particularly frequent in the urban fringes of smaller cities.

Slesnick (1993) points out that the choice of income versus total expenditure is critical for poverty measurement. In the postwar United States, Slesnick finds poverty rates based on total expenditure to be substantially lower than those obtained using income. The author ascribes a possible explanation to consumption smoothing behavior. The permanent income hypothesis (PIH) maintains that consumption

decisions are based only on the permanent component of current income. The households with low income levels may be so due to a large negative transitory component of income. If consumption does not depend on temporary reductions of income as suggested by the PIH, then those households with a temporary low income will show a high consumption to income ratio.

Over time, measures of well-being -- based on a permanent income consumption function and referring to a fixed poverty line -- classify an increasingly higher proportion of households with transitory reduction in income if current income is used. By analogy, it is argued that welfare measures based on consumption should classify households as poor on the basis of their level of total expenditure in order to identify the "permanent income poor." Slesnick finds that the consumption-based poverty rates are uniformly lower than those based on income. Moreover, this distinction permits identifying a set of attributes for the "permanent income poor" in line with prior beliefs. The "consumption poor" have less physical assets such as homes and consumer durables. A large proportion of their total expenditures is spent on necessities such as food and energy. On average, they do not exhibit substantial dis-saving. The evidence described by Slesnick reinforces the important fact that the households most in need are better identified using consumption rather than income as a metric for welfare measurement.

VI. Data Description Using Non-Parametric Density Estimation

Non-parametric analysis is a useful tool to learn from the data about the shape and properties of the conditional expectation function and to derive evidence about the most proper parametric specification. Due to these characteristics, kernel-based techniques are an essential complement to classical methods. Furthermore, the graphical analysis is particularly convenient to summarize the statistical information of large micro-datasets at the exploratory stage of data analysis.

Non-parametric estimation does not assume any functional specification of the conditional expectation $E(y|X)$ or the underlying multivariate distribution $f(y,X)$. This advantage cannot be fully

exploited since the rate of convergence in distribution of non-parametric techniques applied to multivariate density and regression estimation is inversely related to the number of conditioning variables.

In a finite sample it is not possible to estimate the density function at each point. It is however possible to estimate the density f(x) for the set of points belonging to a neighborhood of x. The estimation of a density implies the critical choice of the optimal size of the band centered at x. This decision depends on the sample fraction falling within each interval and the accuracy of the density estimation. Given a bandwidth of h>0, the density estimator can be written as:

$$f(x) = \frac{1}{Nh}\sum_{i=1}^{N} 1\left(-\frac{1}{2} \leq \frac{x-x_i}{h} \leq \frac{1}{2}\right). \qquad \textbf{(14)}$$

Each x in the sample takes the value of 1 if $x_i \in [-1/2, 1/2]$, 0 otherwise. The band width controls the degree of smoothing or regularity of the density curve. Very small values tend to give irregular density estimates, while large values produce very regular estimates[5].

As the bandwidth vary, we generate a discontinuity in the density function. We can circumvent this problem by assigning greater weights between 0 and 1 to the points closer to x. As a consequence, the points close to the band limits have a negligible weight. This can be represented by the following representation of a density:

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right) \qquad \textbf{(15)}$$

where the function K is termed *kernel*. The K(.) weighted function can take different shapes. It must be positive and integrable to 1 within the band limits, symmetric in the neiborhood of zero and decreasing in the absolute value of its argument.

---

[5] The bandwidth tend to 0 as N tends to infinity.

The statistical accuracy of a density estimator $\hat{f}(x)$ can be described by point or global measures. The Mean Square Error (MSE) measures the accuracy of $\hat{f}(x)$ as predictor f (x) at point x:

$$MSE[\hat{f}(x)] = E\left[|\hat{f}(x) - f(x)^2|\right] = Var[\hat{f}(x)] + Bias|[\hat{f}(x)]|^2 \qquad \textbf{(16)}$$

In random sampling from the distribution of x, in general, $\hat{f}(x)$ is a biased estimator of f(x):

$$E[\hat{f}_h(x)] = \frac{1}{h}E\left[K\left(\frac{x-X}{h}\right)\right], \qquad Var[\hat{f}_h(x)] = \frac{1}{nh^2}Var\left[K\left(\frac{x-X}{h}\right)\right] \qquad \textbf{(17)}$$

Given h, the bias of the density estimator does not depend on the sample size and the variance goes to zero as the sample size n rises.

The global measure of statistical accuracy of the density estimator integrates the MSE over x to determine the Mean Integrated Square Error (MISE). This distance between the functions $\hat{f}(x)$ and f(x) describes the degree of concentration of the stochastic function $\hat{f}(x)$ around f(x) as:

$$MISE(\hat{f}_h) = \int_{-\infty}^{\infty} E[|\hat{f}_h - f(x)|^2]dx = \int_{-\infty}^{\infty} MSE[\hat{f}_h(x)]dx. \qquad \textbf{(18)}$$

The choice of the optimal bandwidth for a given sample size corresponds to the minimization of the MISE with respect to h. Given an optimal window, the optimal kernel is a function K^* that minimize the MISE under the non negativity, summability to 1, and zero expected value set of constraints. The Epanechnikov kernel is the one that meets all these conditions:

$$K(x) = \begin{cases} \frac{3}{4}\left(1 - \frac{1}{5}x^2\right) \Big/ \sqrt{5} & if\ |x| < \sqrt{5} \\ 0 & otherwise \end{cases} \qquad \textbf{(19)}$$

The loss of efficiency for not using optimal kernels is in general small. What is crucial is the

choice of the bandwidth controlling the degree of curve smoothing and the trade-off between variance and bias. A large window implies a smooth estimated density with low variance but large bias. On the other hand, a small band implies a variance far from the true value though with less distorted parameter estimates. In general, it is preferable to report density estimates with different degrees of smoothing and let the visual examination evaluate the degree of irregularity of interest.

For this reason, our graphical analysis reports, where computationally feasible, both the Epanechnikov and the Parzen kernel

$$K(x) = \begin{cases} \frac{4}{3} - 8x^2 + 8|x|^3 & if\ |z| \le 1/2 \\ 8(1-|x|)^3/3 & otherwise \end{cases} \qquad \textbf{(20)}$$

that, at same bandwidth, is more sensitive to peaks than the Epanechnikov.

Figures 2, 3 and 4 show the graphs of the Epanechnikov univariate kernel densities for the logarithm of income, expenditure and for the food share over the complete sample.[6] The comparison of the kernels for ln y and ln x indicates that the distributions are both normals. The transformation from incomes to expenditures preserve the means but widens the spread. This evidence implies that there should be statistical loss deriving from the choice of using total expenditure in lieu of income. The food share in Colombia is also distributed as a normal. The sample is more concentrated towards the low end of the distribution. This characteristic is a-typical for a micro data set of a developing country, but it is acceptable if we recall that the sample refers to urban families.

The densities for Food, Adult and Child Clothing reported in Figure 5 are estimated using a 10 percent random subsample stratified by region. As it is apparent from the comparison of the first graph of Figure 5 and Figure 4, the small subsample maintains the information of the full sample. The density

---

[6] The bandwidth has been determined as $h=0.9m/n^{1/5}$ where $m=\min(\sqrt{var_x}, \text{interquartile range}_x/1.349)$ and n = 26,288, the sample size.

estimates for adult and child clothing are close to a Chi-square distribution due to the presence of a high proportion of zero expenditures. It is interesting to note, though, that adult and child clothing are similar goods as far as the distribution shows.

The scatter plots in Figure 6 describe the data clouds for the food, adult and child clothing budget shares plotted against the logarithm of total expenditures and summarized by the budget share curves as simple second order polynomial fits. As expected, the Engel relationship for food is negative and the linear specification seems to interpret correctly the information in the data cloud. For adult and child, clothing the relationship is positive (probably not significant for child clothing) and clearly nonlinear in the raw micro-data. This provides an indication that the proper rank of the demand system, that is the number of Engel curves spanning the function space, should be higher than two and the specification coherent with the data should include higher order income terms.

VI. Conclusion

This study examined some of the issues that are most often encountered when analyzing the socio-economic information of expenditure surveys. Using the Colombian 1985 urban expenditure survey as an example, it discussed the problem of estimating prices when only information on expenditures and demographic characteristics is available, the different options available when aggregating commodities and defining demographic or labor information, and the advantages and drawbacks in choosing total expenditure as a proxy for income or as the metric for poverty measurement. The study also applies non-parametric density estimation as a form of exploratory data analysis and as a means to learn from the data about the most proper parametric specification. It is hoped that this document may simplify the access to the relevant socio-economic information normally provided with the expenditure surveys and increase the researchers' awareness of the critical issues that must be faced when managing data from micro data sets.

References

Barten, A.P. (1964): "Family Composition, Prices and Expenditure Patterns," in Econometric Analysis for National Economic Planning: 16th Symposium of the Colston Society, ed. by P.Hart, G.Mills, and J.K. Whitaker. London, Butterworth, 1964.

Blundell, R., and C. Meghir (1987): "Bivariate Alternatives to the Tobit Model," *Journal of Econometrics*, 34, 179-200.

Cragg, J.G. (1971): "Some Statistical Models for Limited Dependent Variables with Applications to the Demand for Durable Goods," *Econometrica*, 39, 829-844.

DANE (Departemento Administrativo Nacional de Estadistica) (1989): "La Probeza en Colombia," Tomo II, Bogota`, Colombia.

Deaton, A., and M. Irish (1984): "Statistical Models for Zero Expenditures in Household Budgets," *Journal of Public Economics*, 23, 59-80.

Frisch, R. (1959). "A Complete Scheme for Computing All Direct and Gross Demand Elasticities in a Model with Many Sectors." *Econometrica*, 4:1-39.

Glewwe, P. and K. Twum-Baah (1991). "The Distribution of Welfare in Ghana, 1987-88." Working Paper No.75. Living Standards Measurement Study. The World Bank.

Keen, M. (1986): "Zero Expenditures and the Estimation of Engel Curves," *Journal of Applied Econometrics*, 1, 277-286.

Lewbel, A. (1989). "Identification and Estimation of Equivalence Scales under Weak Separability." *Review of Economic Studies*, 56:311-316.

Lee, L.F., and M. Pitt (1986): "Microeconomic Demand Systems with Binding Nonegativity Constraints: the Dual Approach," *Econometrica*, 5, 1237-1242.

Myles, G.D. (1990): *Measurement and Modelling in Economics*, North-Holland.

Perali, C.F. And T.L. Cox (1994): "Demographics, Barten-Prices, Reservation Prices and Quality," Mimeo, University of Wisconsin, Madison.

Pollak, R., and T.J. Wales (1981): "Demographic Variables in Demand Analysis," *Econometrica*, 49, 1533-15.

Pudney, S. (1989): *Modelling Individual Choice*, Basil Blackwell, Oxford.

_____ . (1990): "The Estimation of Engel Curves", in *Measurement and Modelling in Economics*, ed. by G.D. Myles, Elsevier Science Publishers B.V., North-Holland.

Scott C. and B. Amenuvegbe (1990). "Effect of Recall Duration on Reporting of Household Expenditures.

An Experimental Study in Ghana." Working Paper No.6. Social Dimensions of Adjustment in Sub-Saharan Africa. Surveys and Statistics. The World Bank.

Slesnick, D.T. (1993). "Gaining Ground: Poverty in the Postwar United States." *The Journal of Political Economy*, 101:1-38.

Table 1. Distribution of the Sample by City.

Cities	City code	Number of Households
Santa Fé de Bogotá	1	3744
Medellín	2	3744
Cali	3	3120
Barranquilla	4	2496
Bucaramanga	5	1872
Manizales	6	1872
Pasto	7	1872
Cartagena	8	1248
Cúcuta	9	1248
Pereira	10	1248
Ibagué	11	1248
Montería	12	1248
Valledupar	13	1248
Neiva	14	1248
Villavicencio	15	1248
TOTAL		28704

Table 2. Summary Statistics for the Conditional and Unconditional Samples - Shares and Expenditures.

Variable	Min	Max	Mean	Std Dev	Trunc	Min	Max	Mean	Std Dev
WARR	0	0.4534	0.024	0.0265	21.92	0.0001	0.4534	0.0308	0.0262
WMAI	0	0.289	0.0068	0.0129	50.86	4e-05	0.289	0.0139	0.0156
WOCE	0	0.3146	0.0275	0.0267	8.97	0.0001	0.3146	0.0302	0.0265
WYUC	0	0.1231	0.0055	0.0084	38.38	0.0001	0.1231	0.0089	0.0092
WPAP	0	0.3727	0.0151	0.0203	16.99	5e-05	0.3727	0.0182	0.021
WORA	0	0.2326	0.0145	0.0158	15.95	8e-05	0.2326	0.0172	0.0158
WAZU	0	0.4147	0.0217	0.026	21.02	0.0002	0.4147	0.0275	0.0264
WFRI	0	0.2696	0.0086	0.0166	56.00	0.0002	0.2696	0.0195	0.0203
WOLE	0	0.197	0.0041	0.0093	66.39	0.0001	0.197	0.0123	0.0125
WVEG	0	0.3102	0.0254	0.0225	9.25	0.0001	0.3102	0.028	0.0221
WFRU	0	0.24	0.0169	0.0198	24.50	0.0002	0.24	0.0224	0.0199
WCRE	0	0.5486	0.0835	0.064	9.84	0.0001	0.5486	0.0926	0.0608
WOCA	0	0.5476	0.034	0.0348	13.07	0.0003	0.5476	0.0392	0.0345
WPOL	0	0.3409	0.0137	0.0253	61.26	0.0003	0.3409	0.0355	0.0296
WLEC	0	0.4286	0.0462	0.0377	6.46	0.0004	0.4286	0.0494	0.0369
WACE	0	0.3544	0.0226	0.025	28.48	0.0001	0.3544	0.0316	0.0243
WOAL	0	0.4357	0.0279	0.0272	8.76	9e-05	0.4357	0.0305	0.027
WAFH	0	0.8873	0.0146	0.0506	65.96	2e-05	0.8873	0.0429	0.0794
WVIV	0.0002	0.9133	0.1534	0.0799	0.00	0.0002	0.9133	0.1534	0.0799
WSAL	0	0.8239	0.0456	0.0612	9.56	6e-05	0.8239	0.0504	0.0624
WEDU	0	0.7766	0.0873	0.0834	9.77	9e-05	0.7766	0.0967	0.0824
WVEA	0	0.6369	0.073	0.0696	9.81	0.0002	0.6369	0.0809	0.0687
WVEN	0	0.4018	0.0226	0.0329	38.63	0.0002	0.4018	0.0369	0.0352
WENE	0	0.8382	0.0488	0.0913	30.24	0.0001	0.8382	0.07	0.1023
WOTR	0	0.8972	0.1564	0.1403	0.41	0.0007	0.8972	0.1571	0.1403
GTO	1.01e+03	2.63e+06	5.74e+04	7.90e+04	0.00	1.01e+03	2.63e+06	5.74e+04	7.90e+04
WALIM	0.0012	0.994	0.4128	0.1951	0.00	0.0012	0.994	0.4128	0.1951
GAL	1.46e+02	2.26e+05	1.63e+04	1.09e+04	0.00	1.46e+02	2.26e+05	1.63e+04	1.09e+04

Table 3. Summary Statistics for the Conditional and Unconditional Samples - Prices.

Variable	Min	Max	Mean	Std Dev	Trunc	Min	Max	Mean	Std Dev
PARR	0	80	45.06	24.67	21.92	29.92	80	57.7	7.039
PMAI	0	2.40e+02	36.58	46.49	50.86	19.96	2.40e+02	74.43	39.77
POCE	0	7.98e+02	1.68e+02	72.8	8.97	10	7.98e+02	1.84e+02	52.7
PYUC	0	1.00e+02	25.09	24.07	38.38	11.99	1.00e+02	40.72	17.42
PPAP	0	80	20.92	12.45	16.99	4	80	25.21	8.871
PORA	0	1.20e+02	28.12	16.64	15.95	10	1.20e+02	3.35e+01	12.28
PAZU	0	1.40e+02	40.71	22.4	21.02	20	1.40e+02	5.15e+01	8.754
PFRI	0	3.60e+02	79.96	97.69	56.00	40	3.60e+02	1.82e+02	56.53
POLE	0	3.00e+02	50.81	77.49	66.39	25	3.00e+02	1.51e+02	51.84
PVEG	0	8.58e+03	84.99	88.2	9.25	10	8.58e+03	9.37e+01	88.1
PFRU	0	4.23e+02	58.09	47.02	24.50	7.576	4.23e+02	7.69e+01	38.45
PCRE	0	4.40e+02	2.44e+02	94.57	9.84	20	4.40e+02	2.71e+02	51.96
POCA	0	1.00e+03	1.21e+02	1.36e+02	13.07	8.24	1.00e+03	1.39e+02	1.37e+02
PPOL	0	9.00e+02	90.44	1.18e+02	61.26	79.91	9.00e+02	2.33e+02	52.34
PLEC	0	1.02e+03	1.22e+02	1.34e+02	6.46	16.11	1.02e+03	1.31e+02	1.35e+02
PACE	0	1.00e+03	1.55e+02	1.08e+02	28.48	80	1.00e+03	2.17e+02	53.53
POAL	0	7.51e+03	2.71e+02	2.45e+02	8.76	4.545	7.51e+03	2.97e+02	2.41e+02
PAFH	0	1.00e+04	75.34	2.59e+02	65.96	1	1.00e+04	2.21e+02	4.05e+02
PVIV	14.8	4.53e+05	1.54e+03	4.39e+03	0.00	14.81	4.53e+05	1.54e+03	4.39e+03
PSAL	0	9.95e+05	1.20e+03	7.17e+03	9.56	2	9.95e+05	1.32e+03	7.52e+03
PEDU	0	4.69e+05	2.12e+03	6.05e+03	9.77	1.667	4.69e+05	2.35e+03	6.32e+03
PVEA	0	5.17e+04	9.06e+02	1.18e+03	9.81	5	5.17e+04	1.00e+03	1.21e+03
PVEN	0	1.27e+04	3.30e+02	5.02e+02	38.63	3.333	1.27e+04	5.38e+02	5.47e+02
PENE	0	5.42e+05	3.26e+03	1.40e+04	30.24	1	5.42e+05	4.68e+03	1.66e+04
POTR	0	1.20e+06	5.92e+03	2.45e+04	0.41	10	1.20e+06	5.95e+03	2.45e+04

Table 4. Summary Statistics for the Unconditional Sample - Demographics and Income.

Variable	Min	Max	Mean	Std Dev	Variable	Min	Max	Mean	Std Dev
AM0	0	2	0.058	0.239	ETSS	0	1	0.12	0.325
AM1	0	6	0.289	0.555	ETAG	0	1	0.028	0.165
AM2	0	5	0.27	0.542	ETNA	0	1	0.339	0.473
AM3	0	8	0.589	0.895	ENTR	0	1	0.122	0.327
AM4	0	7	0.602	0.753	OOBR	0	1	0.096	0.295
AM5	0	4	0.4	0.513	OEMP	0	1	0.374	0.484
AM6	0	3	0.183	0.397	OEMD	0	1	0.007	0.085
AF0	0	3	0.054	0.231	OIND	0	1	0.34	0.474
AF1	0	5	0.281	0.549	OPAD	0	1	0.06	0.238
AF2	0	4	0.27	0.535	R1	0	1	0.126	0.331
AF3	0	7	0.701	0.921	R2	0	1	0.088	0.283
AF4	0	6	0.752	0.753	R3	0	1	0.139	0.346
AF5	0	5	0.495	0.561	R4	0	1	0.325	0.468
AF6	0	6	0.238	0.473	R5	0	1	0.275	0.446
EA	0	1	0.056	0.23	NPERS	1	25	5.031	2.368
EB	0	1	0.091	0.288	PERCEP	0	9	1.903	1.086
EC	0	1	0.464	0.499	INGJEFE	0	3.08e+06	3.99e+04	6.35e+04
ED	0	1	0.288	0.453	INGHOGA	1.68e+03	9.05e+06	8.00e+04	1.17e+05
EE	0	1	0.1	0.301	PRIMAV	0	1	0.245	0.43
TJ	0	1	0.839	0.367	VERANO	0	1	0.255	0.436
TS	0	1	0.173	0.378	OTONO	0	1	0.25	0.433
EPRO	0	1	0.096	0.295	ARRENDA	0	1	0.33	0.47
EMAN	0	1	0.026	0.158	PROPAGA	0	1	0.424	0.494
EADM	0	1	0.081	0.273	PRONPAG	0	1	0.147	0.354
EMER	0	1	0.188	0.391	INVASOR	0	1	0.099	0.299

Table 5. Frequency Distribution of Family Size.

Value	Count	Cell	Cum
1	553	2.16	2.16
2	2099	8.19	10.34
3	3658	14.26	24.61
4	4768	18.59	43.20
5	4600	17.94	61.14
6	3863	15.06	76.20
7	2303	8.98	85.18
8	1534	5.98	91.16
9	928	3.62	94.78
10	563	2.20	96.98
11	329	1.28	98.26
12	202	0.79	99.05
13	103	0.40	99.45
14	62	0.24	99.69
≥15	79	0.30	100.00
Total	25644		

Figure 1. Frequency Distribution of Family Size.

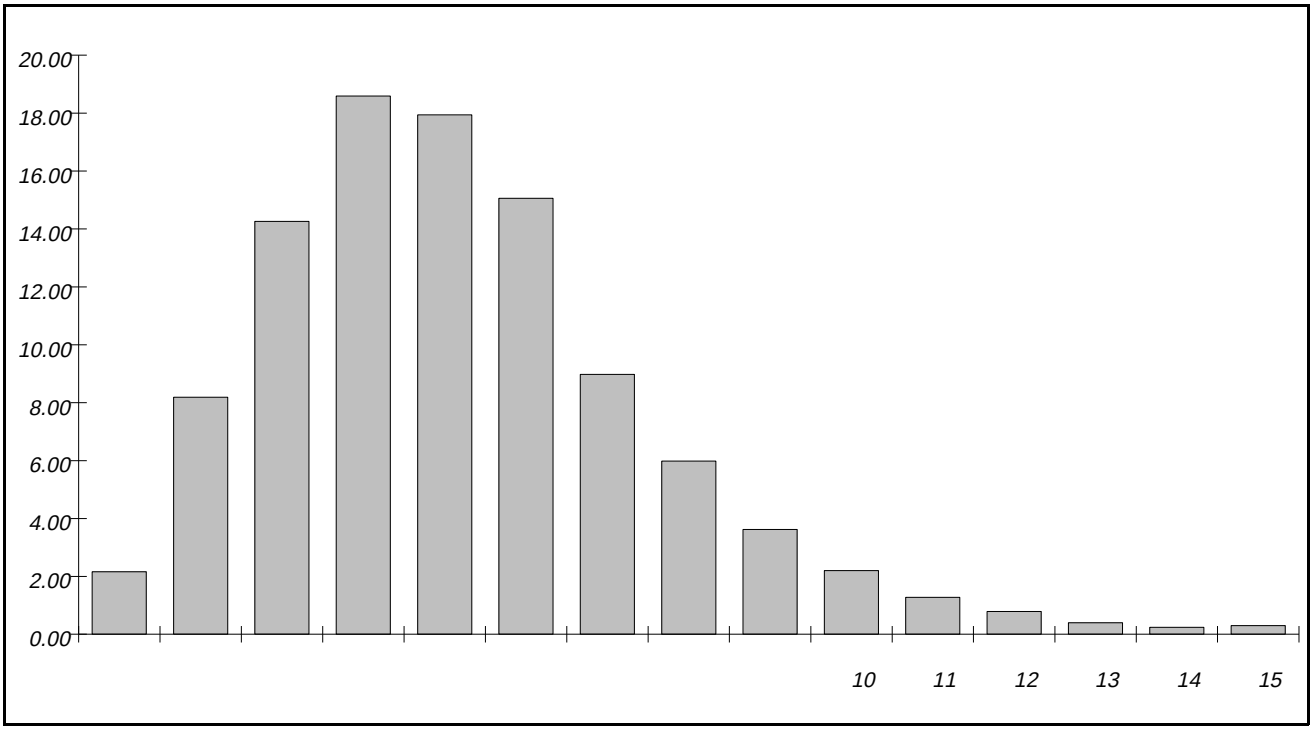


Table 6.a. Percent of Zero Expenditures Across Income Classes for Selected Variables
Total Households: 25,644.

	Income classes					Number of Non Consumers for Each good
	1	2	3	4	5	
	2636	10926	4572	2356	5154	
WARRZ	4.75	29.82	17.74	11.42	36.26	5621
WYUCZ	11.17	35.71	16.52	9.33	27.27	9842
WFRIZ	10.86	40.16	17.30	9.18	22.50	14361
WLECZ	30.90	44.42	10.26	3.62	10.80	1657
WCREZ	13.59	27.43	13.08	9.59	36.31	2523
WAFHZ	12.46	45.81	17.00	8.06	16.67	16915
WSALZ	16.56	47.39	16.76	7.01	12.28	2452
WEDUZ	21.35	51.72	14.72	5.55	6.66	2505
WVEAZ	28.58	47.66	10.61	5.01	8.15	2516
WVENZ	12.50	39.82	16.94	9.16	21.58	9906
WENEZ	21.17	57.29	13.85	4.45	3.24	7755
NCH010	4.21	34.41	20.04	12.03	29.30	9265

Note: The Indigence Line corresponds to the upper limit of class 1; the poverty line to the upper limit of class 2

Table 6.b Percent of Zero Expenditures Within Income Class for Selected Variables
Total Households: 25,644.

	Income classes				
	1	2	3	4	5
	2636	10926	4572	2356	5154
WARRZ	10.13	15.34	21.81	27.25	39.54
WYUCZ	41.69	32.17	35.56	38.96	52.08
WFRIZ	59.14	52.79	54.33	55.98	62.69
WLECZ	19.42	6.74	3.72	2.55	3.47
WCREZ	13.01	6.33	7.22	10.27	17.77
WAFHZ	79.97	70.92	62.90	57.85	54.71
WSALZ	15.40	10.64	8.99	7.30	5.84
WEDUZ	20.30	11.86	8.07	5.90	3.24
WVEAZ	27.28	10.97	5.84	5.35	3.98
WVENZ	46.97	36.11	36.70	38.50	41.48
WENEZ	62.29	40.66	23.49	14.64	4.87
NCH010	14.80	29.18	40.62	47.33	52.68

Note: The Indigence Line corresponds to the upper limit of class 1; the poverty line to the upper limit of class 2

Figure 2. Univariate Kernel Density for ln y.

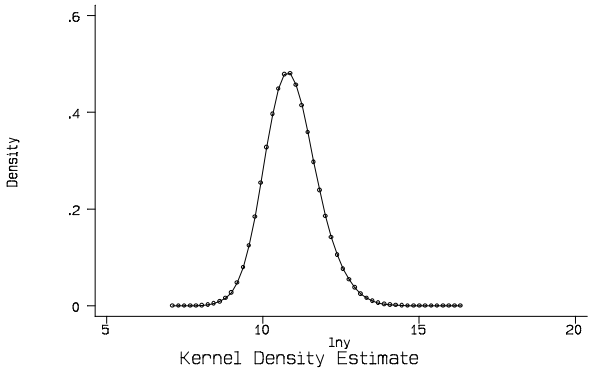


Figure 3. Univariate Kernel Density for ln x

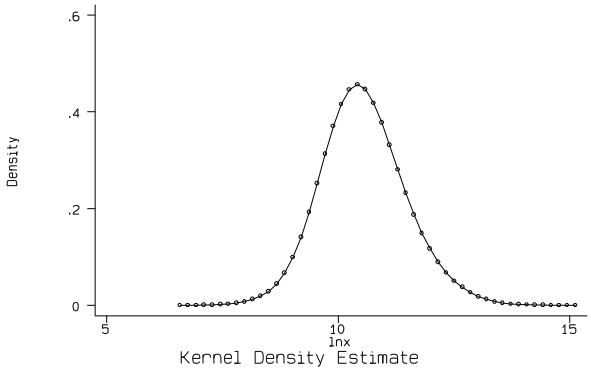
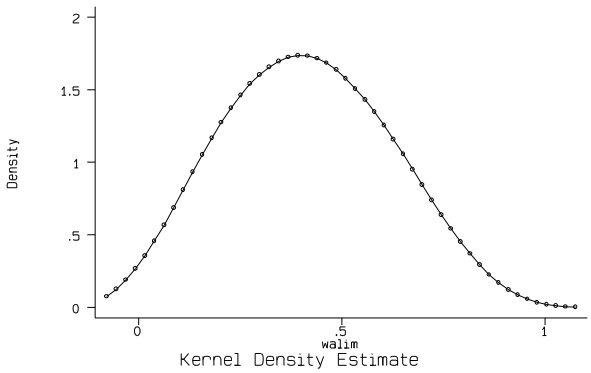


Figure 4. Univariate Kernel Density for the food share



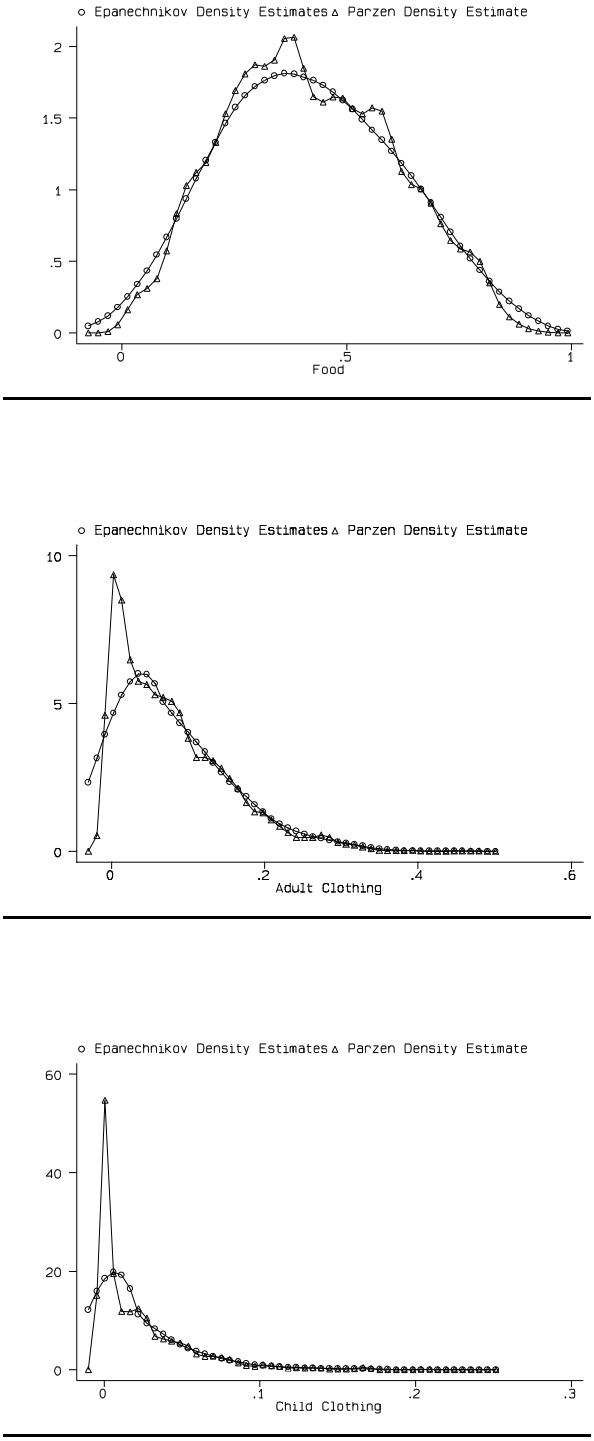
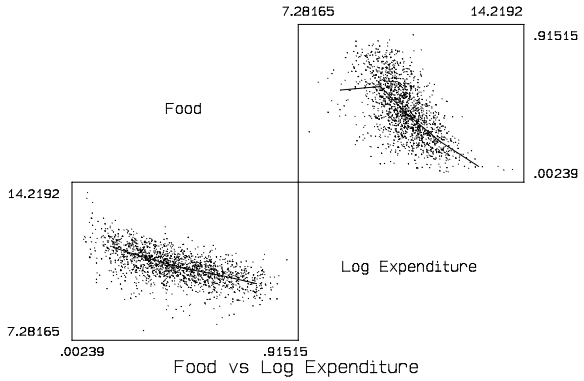
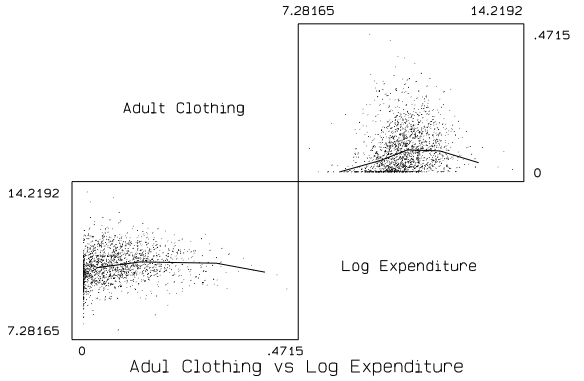


Figure 5. Density estimates for the Food, Adult and Child Clothing on a 10 % random subsample





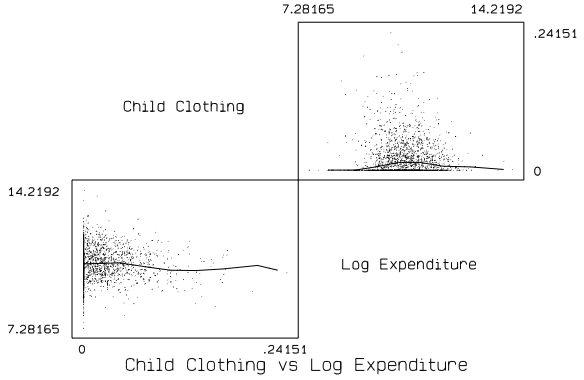


Figure 6. "Engel curve" Scatterplots for Food, Adult and Child Clothing on a 10% random subsample.