Working Paper 95-47
Statistics and Econometrics Series 15
October 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9875

INFLATION AND INEQUALITY BIAS IN THE PRESENCE OF BULK PURCHASES FOR FOOD AND DRINKS

Daniel Peña and Javier Ruiz-Castillo*

Abstract

We study how to improve the estimation of annual food expenditures from household budget surveys with limited information on bulk purchases. We compare three alternative imputation methods by i) estimating the average amount of over or undervaluation according to a Poisson model of the frequency of purchase, ii) measuring the impact on a food share regression of belonging to household types with different bulk purchase habits, and iii) analyzing the outliers exclusively attributable to the shortcomings of each imputation procedure. Finally, we study the implications of different alternatives on the measurement of food price inflation, food expenditure inequality, and household total expenditure inequality.

*Peña, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid; Ruiz-Castillo, Departamento de Economía, Universidad Carlos III de Madrid.

This work is the result of a cooperative agreement with the *Instituto de Estudios Fiscales*. Ana Justel, Coral del Río and Alberto Vaquero provided very able research support. Financial aid by the *Fundación Caja de Madrid*, and from Projects PB93-0230 and PB94-0374 of the Spanish DGICYT is gratefully acknowledged.

INTRODUCTION

The estimation of annual expenditures from information extracted during a limited observation period, poses formidable problems for any household budget survey. In this paper we are concerned with food and drinks for home consumption, or "food expenditures" for short, in the context of the Spanish EPF (*Encuestas de Presupuestos Familiares*), collected by the *Instituto Nacional de Estadística* (INE from now on).

All household members of a certain age are supposed to record all expenditures which take place during a sample week. Then, in depth interviews are conducted to register past expenditures over reference periods beyond a week and up to a year. As far as food expenditures are concerned, in previous surveys all items were assigned a weekly reference period. In recent years, improvements in transportation and storage facilities at home, as well as the rising opportunity cost of time for consumers, have been met on the supply side by improvements in product standarization, package, price and quantity discounts, and a greater availability of both fresh and prepared foods of all types. As a result, bulk purchases have been gaining popularity among certain strata from the more urbanized population. One may conjecture that the habit of acquiring food in large quantities, either regularly or occasionally, by a sizable part of the population, might pose new problems for the estimation of annual food expenditures from the sample information.

We are interested in the impact of different imputation procedures in two areas. In the first place, like in other countries, the INE collects the EPF at regular time intervals in order to estimate the base weights of the official Consumer Price Index system. Thus, a biased estimate of average household expenditures on specific food items, or in the aggregate category as a whole, might lead to a biased estimate of

inflation. In the second place, even if population averages were relatively robust to alternative valuation criteria, biased estimates at the individual level might affect the measurement of household inequality when individual welfare is approximated by total household expenditure.

In the last EPF, which took place during the year from April 1990 to March 1991, the INE collected partial but valuable information on bulk purchases. On the one hand, households were asked to distinguish between minor food expenditures and bulk purchases during the sample week. In both cases, the detailed allocation on specific items was solicited. On the other hand, households were asked whether they had made bulk purchases during the previous three weeks. In these cases the INE only asked for the total amount spent, so that no detailed allocation to specific items was provided.

To decide on the best use of the available information, we need a conceptual framework. Microeconometricians have developed a short but interesting literature on models of consumption behavior using survey data and accounting for the infrequency of purchase question(1). The problem, in Meghir and Robin (1992) words, is that "... observed purchases reflect not only consumption behavior but also purchase policy. Hence, the use of this data to make inferences about consumption requires identifying assumptions linking the observations to the underlying latent variables of interest".

Inspiring ourselves in this work, we suggest a microeconomic behavioral model in which households are assumed to solve their budget allocation problem in two separate stages. Firstly, they decide on the optimal food share, and the allocation of total food expenditure to a set of individual commodities, say the 24 food items for which the INE provides an official monthly consumer price index (plus a residual category on unclassifiable expenditures). This decision is the solution to an underlying

utility maximization problem subject to a budget constraint, in which each household is influenced by a vector of demographic, geographic, socioeconomic, and seasonal characteristics.

Secondly, taking into account the corresponding costs and benefits, households decide whether or not to acquire some of their food and drinks through regular or occasional bulk purchases. As a result of this decision, and taking into account the length of the observation period in the Spanish case, we can classify informally all households into three groups: i) people who make bulk purchases regularly at least once per month, called frequent or F-households; ii) people who make these acquisitions infrequently or occasionally, say every 5, 6, 7 or more weeks, called IO-households; and iii) people who never make a bulk purchase, called N-households.

Under perfect information, the observational consequences of this model are clear. Suppose we start from correct annual data on the share of total expenditures devoted to food. Suppose also that we have a reasonably good regression model of the food share as a function of the set of household characteristics which determine the dependent variable in the cross-section. Then,

i) dummy variables for the F, IO, or N household groups should not be statistically significant, and

ii) outliers in the regression model for the food share should be independent of households purchase policy.

The problem, of course, is that we do not have direct information on the frequency of bulk purchases. What we have is a classification of people into the following 5 groups: households who are never observed to make any bulk purchase (group H1); those observed to have made it only during the sample week (H2), or only during the three weeks prior to the sample week (H3); those observed to have made bulk

purchases in both periods (H4); and a residual contingent (H5) which will be left out of the analysis for reasons to be explained later on.

Notice that all H4-households must be F-households, but that a proportion of the latter are in H3. The rest of group H3, all households in H2, and a proportion of H1 must be IO-households. All N-households are of course in the H1 group. This complex situation precludes an error free imputation of annual food expenditures for every household, and of the aggregate into the 25 food commodities for those lacking detailed information on this matter.

Based on a Poisson model, we start by estimating the household distribution into the F, IO and the N classes, as well as the expected number of bulk purchases in the four week period for the F and IO groups. Therefore, we can compute the average amount that must be added annually per household on account of bulk purchases during the observation period.

We then study the following three alternatives:

(a) Take into consideration only the information from the sample week and, therefore, assign a weekly reference period to all food expenditures during that period -whether they came from small buys or not- but give no weight to bulk aquisitions during the previous three weeks. This is the option chosen by the INE.

(b) Take into consideration all the information from the four week observation period, assigning a weekly reference period only to minor purchases during the sample week, and a 4-week reference period to bulk aquisitions made either during the sample week or prior to it.

(c) Take into consideration all the available information, but also the expected frequency estimated by a Poisson model for the bulk purchases.

The three alternatives are compared by i) estimating the average amount of over or undervaluation according to the Poisson model, ii) inserting dummy variables into the food share regression and measuring the H-effects, and iii) analyzing the outliers exclusively attributable to them.

In practice, since outliers are expected to appear in clusters and masking problems might be present, we use the procedure in Peña and Yohai (1995) that seems to be useful in identifying clusters of outliers avoiding the masking effect. Outliers attributable in each case to erroneous bulk purchase imputations are selected and individually corrected. The three improved versions are compared, and the third one turns out to be favored.

Which are the implications in terms of inflation and inequality bias of maintaining INE's alternative rather than choosing our prefered option? The main conclusions are the following:

i) Because the averaging process taking place in the construction of price indices for the population as a whole, there is little difference between measuring general or food price inflation under the two alternatives.

ii) However, there is a significant improvement in household food expenditure inequality, ranging from 12 to 50 per cent. This range depends on the importance we want to give to economies of scale in consumption, and on which member of the generalized entropy family of indices of relative inequality is used. For the distribution of household total expenditure, the inequality improvement is maintained but, as expected, the range of variation gets drastically reduced up to a 1.5 - 3.0 percent.

The rest of this paper is organized in three sections and an Appendix. Section I presents the data, the notation, the Poisson model for

the frequency of bulk purchase, and the three alternatives. Section II is devoted to the regression analysis of all alternatives, before and after the correction for outliers directly attributable to their known shortcomings. Section III discusses the consequences for the measurement of inflation and inequality of adopting our prefered alternative *versus* the one originally suggested by the INE. The Appendix is devoted to the description of household characteristics, the regression results for the full model, and the allocation of aggregate food expenditures to the 25 specific commodities for those households lacking information on such a breakdown.

I. DATA, NOTATION, AND THE THREE ALTERNATIVES

I.1. The available information

As indicated in the Introduction, the EPF is a household budget survey, collected with the main purpose of providing the weights for the official consumer price index. Because the INE seeks a great geographical detail, EPF samples are usually rather large. Thus, for instance, the latest version for 1990-91 has 21.155 observations for a population of about 11 million households.

Let us denote by BP and SE the *bulk purchases* and *small expenditures* during the sample week, respectively, and by PBP the bulk purchases in the three weeks *prior* to the sample week. We classify all households into five groups as shown in Table 1.

TABLE 1. Household Classification

Variable	Definition	Interpretation
H1	if $SE > 0$, $BP = PBP = 0$	No bulk purchases observed
H2	if $SE > 0$, $BP > 0$, $PBP = 0$	Bulk purchases only during the sample week
H3	if $SE \geq 0$, $BP = 0$, $PBP > 0$	Bulk purchases only during the previous 3 weeks
H4	if $SE > 0$, $BP > 0$, $PBP > 0$	Bulk purchases in both occasions
H5	if either $SE = BP = PBP = 0$, or all food expenditures come from INE's imputation for wages in kind or self-provided consumption	

The sample and population frequencies, where the latter are estimated using the blowing up factors provided by the INE, are given in Table 2.

TABLE 2. Frequency Distributions by Household Type

	Sample distribution		Population distribution	
H1	15.427	72,9	8.203.138	72,6
H2	404	1,9	193.209	1,7
H3	4.848	23,0	2.670.766	23,7
H4	388	1,8	194.249	1,7
H5	88	0,4	37.145	0,3
All	21.155	100,0	11.298.509	100,0

It is important to notice that some households in groups 2 and 4 did not provide the detailed allocation of bulk purchases during the sample week. We denote these groups by H20 and H40, respectively. Then, we denote by H22 and H44 households with full information in groups H2 and H4, respectively. Thus, out of the 404 observations in group H2, only 325 belong to H22, while the remaining 88 belong to H20. Similarly, out of the 388 households in H4, 321 belong to H44 and 77 to H40.

In Table 3 we present two measures of average expenditures for the three observable variables SE, BP and PBP for each of the six H-groups.

TABLE 3. Average weekly food expenditures

	Weekly expenditures			Weekly expenditures per capita		
Group	SE	BP	PBP	SE	BP	PBP
H1	11.431	-	-	3.770	-	-
H20	12.534	3.973	-	3.274	1.106	-
H22	8.904	1.974	-	2.527	576	-
H3	12.503	-	4.809	3.572	-	1.418
H40	9.973	4.765	5.960	2.923	1.444	1.779
H44	8.388	2.362	5.267	2.327	687	1.516
All	11.608	89	1.233	3.681	26	363

Notice the following three facts. In the first place, for both groups which could not remember detail expenditures in their bulk purchases during the sample week, namely groups H20 and H40, their BP approximately doubles that magnitude for the groups with complete information, namely groups H22 and H44, respectively. This might mean that forgetful households tend to think that they spent more in bulk purchases than households who keep good records of it.

In the second place, recall that the vast majority of H3 households are infrequent or occasional bulk purchasers. Therefore, their PBP expenditures could be compared, to a first approximation, with the corresponding magnitude for other households of that type, namely, BP expenditures for H20 and/or H22 hoseholds. Table 3 indicates that the group H3 is much closer on average to group H20. Therefore, we might conjecture that, because of a certain idealization of the past, the bulk purchases in group H3 are also exagerated.

In the third place, notice that groups H40 and H44 have their PBP rather close to each other, contrary to their experience in BP which was examined above. This might be the case because H44 households tend to suffer also from an idealization of the past effect.

I.2. The Poisson model for the frequency of purchase

We do not have information about the household distribution into the F(requent), IO(infrequent or occasional), and N(never) classes defined in the Introduction. In order to obtain an estimate of such distribution, we assume that the number of bulk purchases in a four week period for people in classes F and IO follows a mixed distribution $\alpha_1 P(\lambda_1) + \alpha_2 P(\lambda_2)$, where α_1 and α_2 are the proportion of households in each group, and $P(\lambda_i)$ is a Poisson distribution with parameters λ_1 (> 1) and λ_2 (< 1).

Disregarding all households in group **H5**, we know from Table 2 that:

(i) the proportion of people who did not make bulk purchases in the four week period is 0.7284, so that we can write

$$\alpha_1 e^{-\lambda_1} + \alpha_2 e^{-\lambda_2} + (1 - \alpha_1 - \alpha_2) = 0.7284; \qquad (1)$$

(ii) the proportion of people who did not make bulk purchases in the sample week is

$$\alpha_1 e^{-\lambda_1/4} + \alpha_2 e^{-\lambda_2/4} + (1 - \alpha_1 - \alpha_2) = 0.9656; \qquad (2)$$

(iii) the proportion of people who did not make bulk purchases in the three weks before the sample period is

$$\alpha_1 e^{-3\lambda_1/4} + \alpha_2 e^{-3\lambda_2/4} + (1 - \alpha_1 - \alpha_2) = 0.7456; \qquad (3)$$

(iv) the proportion of people who made some bulk purchases in the sample period is

$$\alpha_1 e^{-\lambda_1/4}(\lambda_1/4) + \alpha_2 e^{-\lambda_2/4}(\lambda_2/4) = 0.0344. \qquad (4)$$

We can solve the system of equations (1) to (4) by a grid search on the four parameter space, or by a nonlinear optimization rutine. An approximate solution (in the least squares sense) to these equations is $\alpha_1 = 0.0353$; $\lambda_1 = 1.7678$; $\alpha_2 = 0.4078$; $\lambda_2 = 0.6121$. According to it, frequent people represents 3.5 per cent of the population with an average time between bulk purchases of 2.26 weeks. For infrequent people (roughly 40 per cent of the population), the average time between bulk purchases is 6.53 weeks. The expected number of bulk purchases in the four week period is given by

$$(0.0353 \times 1.7678 + 0.4078 \times 0.6121) = 0.312.$$

This is in agreement with the observed data in the following sense. We can construct a lower bound for the expected number of bulk purchases in the four week period by simply assuming that all **H3** and **H2** households make one bulk purchase in that period, while all **H4** households make 2. Then: 2 x 0.0173 + 1 x 0.254 + 0 x 0.726 = 0.288.

The above optimization problem is badly conditioned, as usually happens in mixed model estimation in which the strong correlation among the parameters produces a function with more than one local maximum. Fortunately, a wide array of solutions all yield a similar value for this crucial parameter, in the range 0.29 to 0.36. Solutions differ in the assignment of households to the two classes **F** and **IO**, with the corresponding adjustment in the λ parameters. If, for example, α_1 increases, then λ_1 decreases so that the product is approximately mantained. The particular solution already analized seems plausible to us and will be used it in the sequel.

To understand the rest of the model implications, assume for simplicity that the expenditures in each bulk purchase are equal to the mean, $\mu(\mathbf{BP})$. Taking into account that there are 13 periods of four weeks in a year consisting of 52 weeks, the average amount that must be added to each household in a year basis is:

$$13 \times 0.312 \times \mu(\mathbf{BP}) = 4.056\mu(\mathbf{BP}).$$

For individual groups, the estimated Poisson model implies that that we must add $13 \times 1.7678 \times \mu(\mathbf{BP}) = 22.98\mu(\mathbf{BP})$ to 3.53 per cent of F-households, and $13 \times 0.6121 \times \mu(\mathbf{BP}) = 7.96\mu(\mathbf{BP})$ to 40.78 per cent of **IO**-households.

I.3. The three alternatives

Under alternative a, used by the INE, food expenditures are defined as

$$\mathbf{A} = 52\mathbf{SE} + 52\mathbf{BP}$$

Information on **PBP** is ignored, but a weekly reference period is assigned to **BP**. Apparently, the INE was interested in a rough approximation to the average food expenditure per household for the population as a whole. The implicit assumption is that, on average, the infravaluation of **PBP** for **H3** households will be offset by the overvaluation of **BP** for **H2** and **H4** households. However, as the INE is adding μ(**BP**) per 0.034 household, this implies an average of 52 x 0.034 x μ(**BP**) = 1.768μ(**BP**). In other words, alternative **a** is missing more than half of the food expenditure increment attributable to bulk purchases.

As far as different subgroups are concerned, we saw that the Poisson model implies that we need to add 22.98μ(**BP**) to 3.53 percent of the population and 7.96μ(**BP**) to 40.78 per cent of the population, whereas alternative **a** is simply adding 52μ(**BP**) to 3.44 percent of **H2** and **H4** households. In brief, this procedure i) underestimates heavily for the population as a whole, and ii) overestimates for a large amount a small percentage of the population.

Under alternative **b**, only **SE** expenditures are assigned a weekly reference period, while aggregate bulk purchases are assigned a four week reference period. Therefore, annual food expenditures are now

$$\mathbf{B} = 52\mathbf{SE} + (\mathbf{BP} + \mathbf{PBP})13.$$

It is not possible to know *a priori* if this alternative over or underestimates on average on each of the groups. We do know that **B** = **A** for **H1**, **B** < **A** for **H2**, **B** > **A** for **H3**, and **B** can be greater or smaller than **A** for **H4**.

From Tables 2 and 3 we obtain that μ(**PBP**) = 1.876μ(**BP**). Therefore, this procedure is adding on average an additional food expenditure of

$$[0.0171\mu(\mathbf{BP}) + 0.2372\mu(\mathbf{PBP}) + 0.0173(\mu(\mathbf{BP}) + \mu(\mathbf{PBP}))]13 = 6.65\mu(\mathbf{BP}) \quad (5)$$

Thus, alternative **b** overestimates total expenditure by rougly 50 per cent. Of course, as under alternative **a**, the unobserved percentage of infrequent households in group **H1** are necessarily undervalued. With this approach, we are adding $24.39\mu(\mathbf{BP})$ to the 23.72 percent of the population in group **H3**, and $13 \times 2.876\mu(\mathbf{BP}) = 37.39\mu(\mathbf{BP})$ to 1.8 percent of the population in group **H4**. According to the Poisson model, we need to add $7.96\mu(\mathbf{BP})$ to 40.78 percent, and $22.98\mu(\mathbf{BP})$ to 3.53 percent. This suggests that groups **H3** and **H4** are probably overvaluated. Moreover, as these two groups receive all the additions, group **H2** is expected to be undervaluated on average, since no increment is applied to it.

Our third procedure seeks to add an average expenditure to match the expected estimated value. This implies a change in the frequency in (5) such that

$$[0.0171\mu(\mathbf{BP}) + 0.2372\mu(\mathbf{PBP}) + 0.0173(\mu(\mathbf{BP}) + \mu(\mathbf{PBP}))]\,\mathbf{y} = 4.056\mu(\mathbf{BP}).$$

Taking into account that $\mu(\mathbf{PBP}) = 1.876\mu(\mathbf{BP})$, we find that $\mathbf{y} = 7.924$, instead of 13. This means that we are adding an average amount of $22.79\mu(\mathbf{BP})$ to 1.73 per cent of frequent households in **H44**, $7.924\mu(\mathbf{BP})$ to a small group of infrequent households in **H2**, representing 1.71 per cent of the population, and $14.86\mu(\mathbf{BP})$ to **H3** households wich constitute 23.72 percent of the population.

For comparison purposes, in Table 4 we present average weekly expenditures, weekly expenditures *per capita* and the share of total expenditures devoted to food for all groups and the population as a whole under the three options.

TABLE 4. Average weekly food expenditures and mean food share

Group	Weekly expenditures a	b	c	Weekly expenditure per capita a	b	c	Food share a	b	c
H1	11.431	11.431	11.431	3.770	3.770	3.770	0,314	0,314	0,314
H20	28.427	16.507	14.957	7.701	4.380	3.949	0,440	0,308	0,284
H22	16.802	10.878	10.107	4.832	3.103	2.878	0,329	0,244	0,230
H3	12.503	17.312	15.435	3.572	4.991	4.437	0,253	0,319	0,296
H40	29.034	20.698	16.512	9.699	6.146	4.888	0,385	0,317	0,274
H44	17.835	16.017	13.039	5.074	4.530	3.670	0,307	0,286	0,248
All	11.963	12.930	12.414	3.785	4.070	3.919	0,300	0,314	0,307

It can be seen that alternative c produces the smallest variability among the groups. Since weekly food expenditures are not expected to vary much among groups, Table 4 suggests that this alternative is to be prefered. However, this analysis does not take into account other household characteristics and therefore can be very misleading. In the next section we will compare the group means once household differences have been taken into account by regression analysis.

II. REGRESSION ANALYSIS

II.1. First set of results for the three alternatives

Our first task, is to place the previous discussion in a multiple regression setting. Following Deaton et al (1989), we select a flexible functional form for the food share equation. Taking INE's as the reference option, we have

$$\mathbf{SHA} \equiv \mathbf{A/TEA} = \alpha + \beta \ln(\mathbf{PCTE}) + \lambda \ln(\mathbf{HS}) + \Sigma_j \delta_j \mathbf{Nj} + \gamma \mathbf{z} + \varepsilon, \quad (6)$$

where:
- **TEA** is household total expenditure when food expenditure is equal to **A**;
- **HS** is household size;
- **PCTE** ≡ **TEA/HS** is *per capita* household total expenditure;
- **Nj** ≡ **HSj/HS**, and **HSj** is the number of household members in jth's age bracket;
- **z** is a vector of explanatory variables which are identified in the Appendix.

Although (6) can be given a formal interpretation in utility theory, we regard the equation as a convenient representation of the expectation of food patterns conditional on the explanatory variables. The starting point for (6) is Working's (1943) Engel curve study, which linearly relates the share of expenditure on each good to the logarithm of per capita total expenditure. Here the effects of household composition are modeled by the inclusion of the logarithm of household size, ln**HS**, together with the ratios **HSj/HS** to capture the additional effects of composition.

To this model, we add up a set of dummy variables **Hi**, where i = 20, 22, 3, 40 y 44, to capture the effect of belonging to any of these groups relative to the reference group **H1**. For each of the **H** groups, descriptive statistics for selected variables entering the regression analysis are included

in the Appendix. In Table 5 we present the coefficient estimates for the variables we are more interested in (with t-values between brackets), total expenditure elasticities, and a measure of the goodness of fit.

TABLE 5. Summary of regression results for different options

	Option a		Option b		Option c	
INTERCEPT	1.7191	(75.1)	1.8269	(79.5)	1.7999	(79.2)
H3	-0.0201	(-10.9)	0.0580	(31.4)	0.0298	(16.3)
H20	0.1664	(13.6)	0.0121	(1.0)	-0.0158	(-1.3)
H22	0.0524	(8.1)	-0.0453	(-7.1)	-0.0614	(-9.6)
H40	0.1718	(12.4)	0.0969	(7.1)	0.0454	(3.3)
H44	0.0505	(8.1)	0.0284	(4.6)	-0.0164	(-2.6)
lnPCTE	-0.1022	(-61.9)	-0.1097	(-66.3)	-0.1079	(-65.8)
Elasticity	0.6597		0.6504		0.6487	
R^2	0.4054		0.4027		0.4041	
Sample size	21.063		21.067		21.067	

The following comments are in order:

i) The complete model for alternative c appears as Model 1 in the Appendix, where the results are briefly discussed. Detailed results for alternatives a and b are very similar and will be provided upon request. In any case, the goodness of fit for all options is satisfactory for this large cross-section. Heteroskedasticy was much improved by the logarithmic transformation of *per capita* total expenditure.

ii) For the sample as a whole, food is clearly a necessity, with a total expenditure elasticity of approximately 0.65 under all options.

iii) As expected, **H3** households appear undervalued in option **a** which does not give any weight to **PBP**. On the contrary, since **BP** are treated as weekly expenditures, groups 20, 22, 40 and 44 appear very significantly overvalued. Households in **H20** and **H40**, who do not register their allocation of bulk purchases to specific commodities, seem to exaggerate the amount spent on food, a fact already apparent in Table 3. Consequently, they appear as particularly overvalued under option **a**.

iv) With regard to option **b**, as expected **H3** and **H4** households are, on average, overvalued. However, the amount of overvaluation is between one half and one third that of **H20** and **H40** under **a**. Group **H22** is now significantly undervalued. Taking into account Table 3, we conjecture that this type of infrequent households spent less than usual on minor weekly items because they were under the shock of a contemporaneous bulk purchase during this same sample week. Although a similar phenomenon must be present among **H40** households, they are known to have an upward bias in their bulk purchases during the sample week. At any rate, **H40** and **H44** households are overvalued, but about half than under alternative **a**. Finally, note that, as expected, the intercept is larger in **b** than in **a** because the overall underestimation is smaller.

v) Option c values **BP** and **PBP** less than option b. Correspondingly, **H40** households are much less overvalued and **H44** are now slightly below the reference group. Infrequent **H20** households remain essencially insignificant, but with a minus sign, while the **H22** group appear heavily undervalued. Possibly the best feature of this option with respect to option b is that the large group of **H3** hoseholds is now much less overvalued.

II.2. Correction for outliers

We have seen how *a priori* considerations on under and overvaluation caused in each of the three alternatives were confirmed by the regression analysis. Therefore, we have grounds to select those outliers which can be attributed to imperfect imputation of bulk purchases. The aim would be to correct them on an individual basis to reach a second, presumably improved version of each alternative.

However, before proceeding in this direction we must check whether some outliers could be explained by other factors. In particular,

the INE performs imputations to subsidized meals at work, and to meals in the household owned restaurant. We find that 23 negative outliers have a low food share because they have a significant imputation of either of these two types. These observations are not corrected, but taken apart in group **H5** in order not to influence the analysis in the sequel.

Suppose that we fit a multiple regression model to a set of n observations in which there exists a subset of n_0 observations undervaluated, that is, the observed response value at these n_0 points is

$$y_{ob} = y_{real} - k_i,$$

where $k_i > 0$. Assuming that the undervaluation occurs randomly and it is not related to the vector of explanatory variables, it is straightforward to show that the expected effect of these outliers is to bias the intercept by $k^*(n_0/n)$, where $k^* = (\Sigma_i k_i)/n_0$. Therefore, if we fit the regression models given in (6) without the **H** dummy variables, we expect to find in each group outliers with the opposite sign that the sign of the dummy variable in the group (see Table 5 for the latter). Since group **H1** may be undervaluated in the three alternatives, we can assume that large negative outliers in that group are due to the underestimation of bulk purchases.

The search for outliers is carried out by the procedure of Peña and Yohai (1995), that has proved to be able to identify groups of outliers avoiding the masking effect. The outliers are tested with a critical value for te studentized residual of 5. This high value has beeen chosen because (i) it is required to avoid correction for small effects, because as explained before the bias of the intercept may lead to a biased estimation; (ii) outliers due to to awrong imputation for bulk purchases are expected to be large, and (iii) the sample size is large. With this procedure, those outliers attributable to wrong bulk purchase imputations for alternatives **a** and **c**, are shown in Table 6.

TABLE 6. Outliers under different options

	OPTION a		OPTION c	
	(-)	(+)	(-)	(+)
Group				
H1	314	-	421	-
H3	112	-	-	127
H20	-	10	1	-
H22	-	9	7	-
H40	-	6	-	3
H44	-	3	1	-
All	426	28	430	130

The correction of these outliers leads to what we call versions aa, bband cc. A summary of results are presented in Table 7, while the full model for version cc, very similar to the other versions, appears as Model 2 in the Appendix.

TABLE 7. Summary of regression results under different options. The full sample

	Option aa		Option bb		Option cc	
INTERCEPT	1.9028	(87.4)	1.9789	(91.9)	1.9697	(91.7)
H3	-0.0165	(-9.5)	0.0491	(28.6)	0.0241	(14.1)
H20	0.1241	(10.8)	0.0111	(1.0)	-0.0160	(-1.4)
H22	0.0408	(6.7)	-0.0484	(-8.1)	-0.0616	(-10.3)
H40	0.1140	(8.7)	0.0900	(7.0)	0.0368	(2.9)
H44	0.0441	(7.5)	0.0247	(4.2)	-0.0192	(-3.2)
lnPCE	-0.1149	(-73.2)	-0.1196	(-77.1)	-0.1192	(-77.0)
Elasticity	0.6231		0.6234		0.6178	
R^2	0.4604		0.4582		0.4631	
Sample size	21.039		21.040		21.039	

The main implications of this corrections are as follows:

i) The only variable which experiments a change worth noting, is the log of household size which becomes significant under the three options. As expected, goodness of fit are substancially improved, with an R^2 of approximately 0.46 for all alternatives, up from 0.40 before outlier corrections. Also, t values are generally improved.

ii) Total expenditure elasticity for the full sample goes down, approximately, from 0.65 to 0.62 in all alternatives.

iii) In option **aa**, **H3** households appear still significantly undervalued after the treatment of outliers, while all the rest, specially groups **H20** and **H40**, remain seriously overvalued.

iv) In option **bb** we observe a clear improvement of the overvaluation of **H44** and **H3** households. Nevertheless, there remains the large overvaluation of group **H40** and the undervaluation of infrequent households in **H22**.

v) In option **cc** the large group **H3** has improved considerably respect option **bb**, and it is now of the same order of magnitude but opposite sign relative to **aa**. In absolute terms, option **cc** dominates clearly alternative **aa** for **H20**, **H40** and **H44** households, and performs worse only for group **H22** which seems to remain undervalued.

III. IMPLICATIONS

Once we have made the best we could with the available information, it is time to explore the consequences of choosing version cc rather than sticking to INE's option **a**.

III.1. The impact on the measurement of inflation

We have measured the inflation for the food (drinks and tobacco) category during 1993 and 1994 under both alternatives. For that purpose, we have constructed a Laspeyres type price index for the population as a whole (including from **H1** to **H5** households).

Let A^h be the food expenditure of household h under alternative **a**, for example, and let w_i^h be the share of A^h (net of the 25th item of unclassifiable expenditures) devoted to food item i. Let $W = (W_1,...,W_{24})$ be the 24 dimensional vector of population shares, where, for each i, W_i is the weighted mean of the w_i^h's, with weights equal to the A^h's. Then the index we use to compare the price vector p_t with base prices p_0 is

$$P(p_t, p_0, W) = \sum_i W_i (p_{ti}/p_{0i}).$$

Under the current Consumer Price Index system, based in 1992, the INE publishes monthly data for the ratios (p_{ti}/p_{0i}). The vector W under alternative **a** is essentially the vector used in the official system. The construction of such vector under alternative **cc** is described in the Appendix.

The results are as follows. Option **a** yields a food price index of 102.38 and 108.22 for 1993 and 1994, respectively. Option **cc** yields 102.40 and 108.24, a small difference indeed.

On the other hand, notice that the share of household total expenditure devoted to food is 0.2996 and 0.3108 for options a and cc, respectively. Not a large difference either. Therefore, we should not expect great differences in the general price index, covering food and the other eight commodity categories. Indeed, under option a our estimates for the general price index are 105.25 and 110.23 for 1993 and 1994, respectively, while under alternative cc they are 105.24 and 110.22 for those same years.

III.2. The impact on the measurement of inequality

To take into consideration different household needs arising from a different household size, s^h, under alternative a, for example, define adjusted food expenditure by

$$z^h(\Theta) = A^h/(s^h)^\Theta, \ \Theta\in[0,1].$$

We have selected the polar cases $\Theta = 0$ and $\Theta = 1$, corresponding to original household food expenditure, and*per capita* household food expenditure, respectively. We have chosen also the case $\Theta = 0.5$, corresponding to an intermediate view about the importance of economies of scale in consumption within the household.

Because of its good properties(2), we have considered the generalized entropy family of relative inequality indices:

$$I_c(\mathbf{z}) = (1/H)[1/c(c-1)][\Sigma_h(z^h/\mu(\mathbf{z}))^c - 1], \qquad c \neq 1, 0;$$

$$I_c(\mathbf{z}) = (1/H)[\Sigma_h(z^h/\mu(\mathbf{z}))\ \ln(z^h/\mu(\mathbf{z}))], \qquad c = 1;$$

$$I_c(\mathbf{z}) = (1/H)[\Sigma_h \ln(\mu(\mathbf{z})/z^h)], \qquad c = 0,$$

where μ is the function providing the distribution mean. In particular, we have selected a member of this family more sensitive to the upper part of the distribution, c = 2 -which is 1/2 the coefficient of variation- and a member more sensitive to the lower part, c = -1. We have estimated also

the two indices originally suggested by Theil corresponding to c = 1 and c = 0.

The results are in the left hand side of Table 8. We observe a systematic improvement in food expenditure inequality with option cc for all values of Θ and all members of the generalized entropy family. The estimated reduction of inequality ranges from a minimum of 12 percent to a maximum of 50 percent. Such an improvement is greater the more sensitive one is to the upper tail of the distribution, and at an intermediate value of the parameter representing the importance of economies of scale.

Finally, we have carried on the same exercise for the distribution of total expenditure. The results are in the right hand side of Table 8. The improvement in inequality persists in this domain, but loses importance: the range of variation is from 1.5 to 3.0 percent.

TABLE 8. Inequality under different options

	Food expenditure inequality				Total expenditure inequality			
Θ = 0.0	c = 2	c = 1	c = 0	c = -1	c = 2	c = 1	c = 0	c = -1
Option a	0.1813	0.1636	0.1853	0.3163	0.2525	0.2046	o.2169	0.3089
Option cc	0.1613	0.1463	0.1593	0.2185	0.2474	0.2021	0.2134	0.2994
a/cc	1.1240	1.1182	1.1632	1.4476	1.0206	1.0123	1.0164	1.0317
Θ = 0.5								
Option a	0.1412	0.1249	0.1341	0.1982	0.2128	0.1701	0.1697	0.2111
Option cc	0.1208	0.1066	0.1089	0.1308	0.2094	0.1674	0.1664	0.2043
a/cc	1.1689	1.1717	1.2314	1.5145	1.0162	1.0161	1.0198	1.0333

$\Theta = 1.0$

Option a	0.1726	0.1414	0.1423	0.1887	0.2575	0.1922	0.1831	0.2179
Option cc	0.1497	0.1224	0.1184	0.1349	0.2535	0.1894	0.1800	0.2123
a/cc	1.1530	1.1552	1.2018	1.3988	1.0158	1.0148	1.0172	1.0264

NOTES

(1) See Pudney (1987) and Meghir and Robin (1992), and the references quoted there.

(2) For a characterization, see for instance Shorrocks (1980). For a defense, discussion and applications, see Cowell (1984), Coulter *et al* (1992a, 1992b), and Ruiz-Castillo (1995).

REFERENCES

Coulter, F., F. Cowell and S. Jenkins (1992a), "Differences in Needs and Assessment of Income Distributions," *Bulletin of Economic Research*, 44: 77-124.

Coulter, F., F. Cowell and S. Jenkins (1992b), "Equivalence Scale Relativities and the Extent of Inequality and Poverty," *Economic Journal*, 102: 1067-1082.

Cowell, F. (1984), "The Structure of American Income Inequality," *Review of Income and Wealth*, 30: 351-375.

Deaton, A., J. Ruiz-Castillo and D. Thomas (1989), "The Influence of Household Composition on Household Expenditure Patterns: Theory and Spanish Evidence," *Journal of Political Economy*, 97: 179-200.

Meghir, C. and J. M. Robin (1992), "Frequency of Purchase and the Estimation of Demand Systems",*Journal of Econometrics*, 53: 53-86

Peña, D. and V. Yohai (1995), "The Detection of Influential Subsets in Linear Regression by Using an Influence Matrix", *Journal of the Royal Statistical Society, Serie B.* 57: 1-12.

Pudney, S. (1987), "On the Estimation of Engel Curves", presented at a *Conference on Measurement and Modelling in Economics*, Nuffield College.

Ruiz-Castillo, J. (1995), "The Anatomy of Money and Real Inequality in Spain, 1973-74 to 1980-81", forthcoming in *Journal of Income Distribution*, **4**.

Shorrocks, A. (1980), "The Class of Additively Decomposable Inequality Measurements," *Econometrica*, **48**: 613-625.

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

APPENDIX

I. VARIABLES DEFINITION

Demographic

HS = household size

Nj = HSj/HS, where HS1 = number of household members less than 4 years old
HS2 = number of household members between 4 and 8 years old
HS3 = number of household members between 9 and 14 years old
HS4 = number of household members between 15 and 17 years old
HS5 = number of household members between 18 and 24 years old
HS6 = number of household members between 25 and 40 years old
HS7 = number of household members between 41 and 64 years old
HS8 = number of household members between 65 and 75 years old
HS9 = number of household members older than 75 years

Socioeconomic

NEARN = number of income earners in the household

S = female household head

HHED1 = household head educational level: illiterate
HHED2 * = without formal studies or only first grade
HHED3 = second grade
HHED4 = high school
HHED5 = three year college degree
HHED6 = other college degrees and graduate studies

SED0 = no espouse
SED1 * = espouse educational level: illiterate, without formal studies, first and second grade
SED2 = high school
SED3 = college degree and graduate studies

SOCIO1 = agrarian working class, and small landowners
SOCIO2 * = non-agricultural working class and other unclassifiable members of the labor force
SOCIO3 = agrarian entrepeneurs, armed forces, non-agrarian entrep. without salaried workers
SOCIO4 = middle and upper class
SOCIO5 = not in the labor force

MIGR = recently inmigrated household head

Housing conditions

SQM = housing living space in square meters

TEN1 * = owner-occupied housing
TEN2 = market rental housing
TEN3 = subsidized public housing
TEN4 = rental housing, unknown legal condition
TEN5 = other housing tenure

BUILD1 * = detached, single housing unit
BUILD2 = building with two housing units

BUILD3 = building with three or more housing units
BUILD4 = non-residencial building

NSRY = number of secondary living quarters

Geographic and seasonal conditions

MUN1 = municipality size: up to 2.000 inhabitants
MUN2 = from 2.000 to 5.000 inhabitants
MUN3 = from 5.000 to 10.000 inhabitants
MUN4 = from 10.000 to 20.000 inhabitants
MUN5 = from 20.000 to 50.000 inhabitants
MUN6 = from 50.000 to 100.000 inhabitants
MUN7* = greater than 100.000 inhabitants

CCAA1 * = Andalucía
CCAA2 * = Aragón
CCAA3 = Asturias
CCAA4 = Baleares
CCAA5* = Canarias
CCAA6 * = Cantabria
CCAA7 = Castilla y León
CCAA8 = Castilla-La Mancha
CCAA9 * = Cataluña
CCAA10 = Comunidad Valenciana
CCAA11 = Extremadura
CCAA12 = Galicia
CCAA13 = Madrid
CCAA14 * = Murcia
CCAA15 = Navarra
CCAA16 = País Vasco
CCAA17 * = La Rioja
CCAA18 * = Ceuta
CCAA19 * = Melilla

SPRING * 1990 = quarter in which the interview took place
WINTER 1991
SUMMER 1991
AUTUMN 1991

WEEK2 = the interview took place during the first two weeks of the month
WEEK4 = the interview took place during the third or fourth week of the month
WEEK5 = the interview took place during the fifth week of the month

NOTE: Dummy variables excluded from the regression are denoted by the symbol *

II. DESCRIPTIVE STATISTICS

Mean of selected continous variables

	H1	H2	H3	H4	All
TE	2.198.608	2.704.966	3.137.648	3.227.491	2.447.747
HS	3,27	3,84	3,77	3,83	3,41
PCTE	737.321	766.088	907.804	936.648	781.685
SQM	102.0	100.2	107.2	107.7	103.3

Percentage distributions of selected discrete variables

NSRY					
0	89,9	90,4	86,3	89,9	89,1
1	9,8	9,0	13,1	10,1	10,5
2 or more	0,3	0,6	0,6	---	0,4
	100,0	100,0	100,0	100,0	100,0
NEARN					
0	0,06	-	-	-	0,04
1	43,3	40,4	38,1	35,3	41,9
2 or more	56,64	59,6	61,9	64,7	58,06
	100,00	100,0	100,0	100,0	100,00
HHED					
1	5,4	2,3	1,7	2,4	4,4
2	63,9	50,3	49,0	43,8	59,7
3	15,2	20,7	19,1	18,0	16,3
4	8,4	16,4	15,3	18,7	10,4
5	3,8	5,9	6,8	7,4	4,6
6	3,3	4,4	8,1	9,7	4,6
	100,0	100,0	100,0	100,0	100,0
SOCIO					
1	7,8	8,4	5,0	3,4	7,1
2	21,3	27,7	26,4	25,3	22,7
3	20,5	24,5	28,8	28,3	22,6
4	8,7	13,1	15,6	21,1	10,6
5	41,7	26,3	24,2	21,9	37,0
	100,0	100,0	100,0	100,0	100,0
MUN					
1	8,2	7,0	4,6	3,9	7,3
2	9,7	6,2	5,8	5,0	8,6
3	11,6	7,6	8,5	6,2	10,7
4	10,9	9,9	8,9	8,3	10,4
5	12,2	7,2	10,5	7,6	11,6
6	8,7	9,6	9,7	7,7	9,0
7	38,7	52,5	52,0	61,3	42,4
	100,0	100,0	100,0	100,0	100,0

III. REGRESSION RESULTS

MODEL 1. Dependent variable: food share under alternative c

INTERCEPT	1.7999	(79.2)	SQM	-0.0001	(-6.9)
H3	0.0298	(16.3)	TEN2	0.0343	(11.1)
H20	-0.0158	(-1.3)	TEN3	0.0445	(11.4)
H22	-0.0614	(-9.6)	TEN4	0.0417	(11.8)
H40	0.0454	(3.3)	TEN5	0.0093	(3.1)
H44	-0.0164	(-2.6)	BUILD2	-0.0133	(-3.6)
lnPCTE	-0.1079	(-65.8)	BUILD3	-0.0136	(-6.4)
lnHS	-0.0026	(-0.8)	NSRY	-0.0284	(-12.2)
N1	-0.0327	(-3.1)	MUN1	0.0269	(7.4)
N2	-0.0509	(-5.6)	MUN2	0.0159	(5.0)
N3	-0.0342	(-4.2)	MUN3	0.0105	(3.7)
N4	-0.0719	(-7.4)	MUN4	0.0091	(3.3)
N5	-0.0797	(-12.3)	MUN5	0.0125	(4.9)
N6	-0.0466	(-9.6)	MUN6	0.0076	(2.8)
N7	0.0071	(1.7)	CCAA3	-0.0096	(-2.2)
N8	0.0132	(3.3)	CCAA4	-0.0096	(-7.2)
NEARN	-0.0057	(-7.4)	CCAA7	-0.0396	(-2.0)
S	0.0089	(3.0)	CCAA8	-0.0064	(-4.4)
HHED1	0.0142	(3.8)	CCAA10	-0.0164	(-6.4)
HHED3	-0.0041	(-1.8)	CCAA11	-0.0358	(-7.9)
HHED4	-0.0156	(-5.5)	CCAA12	0.0306	(10.0)
HHED5	-0.0196	(-4.8)	CCAA13	-0.0193	(-7.8)
HHED6	-0.0245	(-5.3)	CCAA15	-0.0269	(-4.1)
SED0	-0.0238	(-7.6)	CCAA16	-0.0107	(-3.1)
SED2	-0.0060	(-1.8)	WINTER	-0.0070	(-3.3)
SED3	-0.0077	(-4.2)	SUMMER	0.0041	(2.0)
SOCIO1	0.0109	(3.2)	AUTUMN	-0.0096	(-4.6)
SOCIO3	-0.0062	(-2.8)	WEEK2	-0.0026	(-1.7)
SOCIO4	-0.0077	(-2.4)	WEEK3	0.0075	(2.7)
SOCIO5	0.0141	(5.5)			
MIGR	0.0083	(2.3)			
R^2	0.4041				
Sample size	21.067				

All variables with at least a 1.70 t-value in absolute terms in Model 1 which were present in all **H**-groups, were selected for the regression analysis. Demographic composition effects show that, relative to the oldest groups, the presence of younger members has a negative impact on the food share. The number of income earners causes also a significant negative effect. The household head educational variables indicate that the greater the educational level attained, the smaller the food share. The effect of the spouse's educational level, whenever present, is less clear. Higher socioeconomic classes and not beeing a recent inmigrant have significantly smaller food shares. Households enjoying larger housing space, in owner-occupied housing, and in buildings with two or more housing units, have a smaller food share. The smaller the municipality size, the greater the expenditure devoted to food. Only relatively poor and agrarian Galicia has a greater food share than Andalucía. Aragón, Cantabria, Canarias, Cataluña, Murcia, La Rioja, and the North-African cities Ceuta and Melilla, have no significant effect. The quarter or/and the week in which the survey took place does not cause a clearly interpretable pattern.

MODEL 2. Dependent variable: food share under alternative cc

INTERCEPT	1.9697	(91.7)	SQM	-0.0001	(-7.7)
H3	0.0241	(14.1)	TEN2	0.0348	(12.1)
H20	-0.0160	(-1.4)	TEN3	0.0427	(11.7)
H22	-0.0614	(-10.3)	TEN4	0.0410	(12.4)
H40	0.0368	(2.9)	TEN5	0.0109	(4.0)
H44	-0.0192	(-3.3)	BUILD2	-0.0151	(-4.4)
lnPCTE	-0.1192	(-77.0)	BUILD3	-0.0140	(-7.0)
lnHS	-0.0145	(-4.6)	NSRY	-0.0265	(-12.2)
N1	-0.0306	(-3.1)	MUN1	0.0148	(5.0)
N3	-0.0302	(-3.9)	MUN3	0.0119	(4.4)
N4	-0.0706	(-7.8)	MUN4	0.0083	(3.2)
N5	-0.0755	(-12.5)	MUN5	0.0122	(5.1)
N6	-0.0521	(-11.5)	MUN6	0.0102	(3.9)
N7	0.0047	(1.2)	CCAA3	-0.0099	(-2.4)
N8	0.0109	(2.9)	CCAA4	-0.0376	(-7.3)
NEARN	-0.0049	(-4.9)	CCAA7	-0.0073	(-2.5)
S	0.0028	(1.0)	CCAA8	-0.0185	(-5.3)
HHED1	0.0165	(4.7)	CCAA10	-0.0178	(-7.3)
HHED3	-0.0043	(-2.0)	CCAA11	-0.0432	(-10.3)
HHED4	-0.0134	(-5.1)	CCAA12	0.0337	(11.8)
HHED5	-0.0201	(-5.3)	CCAA13	-0.0187	(-8.1)
HHED6	-0.0243	(-5.7)	CCAA15	-0.0265	(-4.3)
SED0	-0.0175	(-6.0)	CCAA16	-0.0102	(-3.2)
SED2	-0.0020	(-0.6)	WINTER	-0.0070	(-3.6)
SED3	-0.0140	(-3.5)	SUMMER	0.0042	(2.2)
SOCIO1	0.0103	(3.2)	AUTUMN	-0.0096	(-4.6)
SOCIO3	-0.0058	(-2.8)	WEEK2	-0.0026	(-1.8)
SOCIO4	-0.0047	(-1.5)	WEEK3	0.0062	(2.4)
SOCIO5	0.0131	(5.5)			
MIGR	0.0081	(2.4)			
R^2	0.4631				
Sample size	21.039				

The most important difference is in the coefficient of the log of household size, lnHS, which is now clearly significant and it was not before. Not having a spouse, or having one highly educated, depresses the food share. All other patterns present in Model 1 are maintained, although four variables -N7, S, SED2 and SOCIO4- are no longer significant.

IV. ALLOCATION

In option **c** we have made the best possible imputation of annual food expenditures from the available information during a four week observation period. However, for **H20** and **H40** households bulk purchases made during the sample week must be allocated among the 25 specific food items. The same must be done for bulk purchses during the prior three weeks for **H3** , **H40** and **H44** households.

We start from the hypothesis that people might not buy goods in the same proportion in a bulk purchase, possibly in a large discount store in a shopping mall, than in smaller acquisitions during weekly errands in their neighbourhood. We have complete information in this respect for **H22** and **H44** households. Based on the shopping behavior of these groups, we have classified 25 commodities into bulk purchase-goods, weekly-goods, and other-goods. For every i = 1,..., 25, let us denote by $\mathbf{BPW_i}$ and $\mathbf{SEW_i}$ the share of **BP** and **SE** expenditures, respectively, devoted to good i. Whenever the variable ($\mathbf{BPW_i}$ - $\mathbf{SEW_i}$) takes a sizable positive value for both **H22** and **H44** households, we say that good i is a bulk purchase-good. Whenever it takes a negative value for both groups, we say that it is a weekly-good. If this variable takes small values and/or different signs depending on the group, then we classify it as an other-good.

Following this criterion, we partition the set into 9 bulk purchase-goods, 8 weekly-goods, and 8 other-goods. This is a reasonable classification: i) prepared goods of all sorts appear prominently in bulk purchases; ii) all types of fresh items appear as weekly goods; iii) different meats, milk, both alcoholic and non-alcoholic drinks, as well as tobacco which is only bought in special stores, appear as neither and form a group of its own.

In the next step, before deciding on an allocation procedure for the above household groups, we would like to learn as much as possible about their behavior in this 25-dimensional commodity space. Of course, at this level of detail, for households in groups **H1**, **H3**, **H20**, and **H40** we have only information on **SE** expenditures. Nevertheless, we run two types of regressions for the sample of 21.039 observations remaining after the outliers analysis leading to option **α**. In the first place, we run 25 regressions to compute total expenditure elasticities for each good. These are presented as column (1) in Table A. In the second place, we run 25 regressions to explain the allocation of aggregate food expenditure under alternative **c** to the 25 food commodities. Per thousand commodity shares, as a proportion of aggregate food expenditures, are presented in column (2) in Table 10. Regression coefficients for the 5 groups, relative to the **H1** reference group, are presented in columns (3) to (7). Non significant coefficients are singled out by means of an arterisk. Finally, each equation's R^2 is provided in column (8).

i) We are mostly interested in learning as much as possible about the largest of all difficult groups, namely, **H3** households. These households, who were observed to make some bulk purchase only during the three weeks prior to the sample week, contain a large proportion of people who make a bulk purchase every four weeks or more. Given the above classification, we expect them to be short of bulk purchase-goods, long on weekly-goods, and close to the reference group in other-goods. Non counting tobacco, **H3** households satisfy the expected pattern in 13 cases, present a single violation in other-goods, and non significant coefficients in the remaining 10 cases.

ii) It is illuminating to compare this evidence with the case of infrequent or occasional bulk purchasers who made their large acquisitions during the sample week. In only two bulk purchase-goods, one weekly-good and one other-good **H22** households differ from the reference group.

iii) Groups **H20** and **H40** do not provide information on their bulk purchase commodity breakdown. Their allocation of **SE** expenditures should not be very different

from the reference group. In any case, they should resemble H3 households in being short on bulk purchase-goods and long on weekly-goods. The result is that, not counting tobacco, group H20 differs from H1 only in 6 occasions, and H40 in 5. In 9 out of these 11 cases, they behave as expected.

iv) If the behavior of frequent bulk purchasers in H44 had been well captured by the regression model, their dummy variables should not be significant. This is indeed the case in all but two cases: milk and alcoholic drinks, to which they devote a smaller and a greater share of food expenditures, respectively.

TABLE A. Results for individual commodities

GOODS:	(1) Total exp. elasticity	(2) Comm. share	(3) H3	(4) H20	(5) H22	(6) H40	(7) H44	(8) R^2
Bulk purchase								
1. Oils	0,709	35,4	-0.0055	-0.0195	*	*	*	0,0507
2. Prep. fish	1,010	37,8	*	*	0.0093	-0.0177	*	0,0450
3. Prep. vegts.	0,672	19,7	-0.0025	*	*	*	*	0,0203
4. Other foods	0,916	27,1	*	*	*	*	*	0,0353
5. Coffee, tea, cocoa, etc.	0,729	13,9	-0,0023	*	*	*	*	0,0390
6. Other meats	0,750	92,7	*	-0.0245	*	*	*	0,0643
7. Milk prods.	0,718	43,2	-0.0025	*	*	-0.0199	*	0,0336
8. Sugar	0,366	6,6	-0.0018	-0.0039	-0.0022	-0.0045	*	0,0587
9. Fruit preserves	0,786	9,5	*	*	*	*	*	0,0171
Weekly								
10. Bread	0,096	65,2	0.0037	*	*	0.0185	*	0,3284
11. Fresh vegts.	0,568	45,1	0.0020	*	*	*	*	0,0883
12. Potatoes	0,439	18,0	*	*	*	*	*	0,0972
13. Fresh fruit	0,575	81,3	*	*	-0.0119	0.0628	*	0,1023
14. Eggs	0,343	18,8	*	*	*	*	*	0,0413
15. Fresh and frozen fish	0,752	69,2	*	*	*	*	*	0,0874
16. Unclassi-fiable	1,406	24,4	*	0.0526	*	*	*	0,0420
17. Grains	0,780	57,3	*	*	*	*	*	0,0441
Other								
18. Beef	0,846	62,1	*	0.0297	*	*	*	0,1500
19. Lamb	0,932	22,5	*	*	*	*	*	0,0729
20. Pork	0,562	31,5	*	*	*	*	*	0,0744
21. Chicken	0,394	43,1	*	*	*	*	*	0,0411
22. Milk	0,344	68,3	-0.0052	-0.0246	-0.0094	*	-0.0093	0,1065
23. Non-alc. drinks	0,820	19,6	*	*	*	*	*	0,0510
24. Alcoholic drinks	0,980	31,2	*	*	0.0111	*	0.0088	0,0423
25. Tobacco	0,593	56,4	0.0064	0.0286	*	0.0415	*	0,1547

The main thrust of this analysis is that H-groups behave in the 25 commodity space in general agreement with our expectations based on evidence from their aggregate food behavior. This is helpful in solving our allocation problem in this commodity space. For all households involved, our criterion is to allocate those totals into the 25 items according to the population means. Essencially, in this way we correct **H3, H20** and **H40** households in an appropiate direction: given that they incurred in BP or PBP but we do not have any detailed breakdown, we raise their share of bulk purchase-goods, and lower their share of weekly-goods.