The Estimation of Food Expenditures From Household Budget Data in the Presence of Bulk Purchases

Daniel PEÑA
Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Madrid 126, 28903 GETAFE, Spain (dpena@est-econ.uc3m.es)

Javier RUIZ-CASTILLO
Department of Economics, Universidad Carlos III de Madrid, Madrid 126, 28903 GETAFE, Spain (jrc@eco.uc3m.es)

The aim of this article is the estimation of annual food expenditures with limited information about bulk purchases with data from a Spanish household-budget survey for 1990–1991. Three alternatives are compared. The first, currently used for official purposes, does not use all the information. The second uses all the available information in a rough way. The third assumes a formal model for the unknown frequency of purchases. The three alternatives are compared by a regression model that should be homogeneous with respect to the dummy variables that represent the partial information of the groups and should show a distinct pattern of outliers under each alternative. Finally, we study the effect of the official and the best alternative on food inflation and inequality measures. We find that they lead to similar inflation rates but to different inequality estimates.

KEY WORDS: Food-expenditure estimation; Outliers; Poisson distribution; Regression.

The estimation of annual expenditures from information extracted during a limited observation period poses formidable problems for any household-budget survey. In the case of food and drink for home consumption, or "food expenditures" for short, there are two purchase modes. On the one hand, many consumers acquire all or part of their food in relatively small quantities once or several times per week. On the other hand, 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 standardization; 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. The juxtaposition of these two purchase modes makes data collection and expenditure estimation difficult tasks for official statisticians.

In this article we are concerned with these issues 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 that 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. In previous surveys from 1958 to 1980–1981, the INE assigned a weekly reference period to all food expenditures. Therefore, annual food expenditures were estimated by multiplying recorded food expenditures by 52. In the last EPF, however, which took place 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, although not all households were able to comply with such detail. 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. The problem we study in this article is how best to use the new information on bulk purchases to estimate each household's annual food expenditures in the 1990–1991 EPF.

Two solutions can be immediately suggested. (a) Take into consideration only the information from the sample week, and 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 acquisitions during the previous three weeks. This is the option actually chosen by the INE. (b) At the other extreme, 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 four-week reference period to bulk acquisitions made either during the sample week or prior to it.

A third solution is to set up a bulk-purchase model and estimate the food expenditure making use of inferences from it. Taking into account the length of the observation period in the Spanish case, we can classify all households informally into three groups—(1) people who make bulk purchases regularly at least once per month, called frequent or F households; (2) people who make these acquisitions infrequently or occasionally, say every five, six, seven, or

Journal of Business & Economic Statistics
July 1998, Vol. 16, No. 3

more weeks, called **I** households; and (3) people who never make a bulk purchase, called **N** households. The problem is that, unfortunately, the INE did not collect information on the frequency with which households make bulk purchases. What we have is a classification of people into the following four groups: (1) households who are never observed to make any bulk purchase (group **H1**), (2) those observed to have made a bulk purchase only during the sample week (**H2**), (3) those observed to have made a bulk purchase only during the three weeks prior to the sample week (**H3**), and (4) those observed to have made bulk purchases in both periods (**H4**). Using this information and assuming that the distribution of purchases in each group **F**, **I**, and **N** follows a Poisson model, we suggest an alternative (c) in which, as in option (b), all available information is used but the reference period for bulk purchases is modified so that, on average, we add an amount per household on account of these purchases equal to the amount expected from the Poisson model.

The three alternatives are compared from two complementary perspectives. In the first place, we estimate the average amount of overvaluation or undervaluation imputed to each **H** group by each alternative. We show that, from this viewpoint, alternative (c) is to be preferred. In the second place, as shown by Meghir and Robin (1992), households are assumed to solve their budget-allocation problem in two separate stages. First, they decide on the optimal food share and the allocation of total food expenditure among a set of individual commodities. Second, households decide whether or not to acquire some of their food and drink through regular or occasional bulk purchases. Under perfect information, the observational consequences of this model are clear. Suppose we have an accurate estimate of each household share of total expenditures devoted to food and a reasonably good regression model of the food share as a function of a wide set of household characteristics. Then, (1) dummy variables for the **H1** to **H4** categories should not be statistically significant and (2) outliers in the regression model for the food share should be independent of households' purchase policy. In the absence of information on the frequency of bulk purchases, options (a), (b), and (c) can be seen as providing alternative assumptions about bulk-purchases reference periods for households in the **H1**–**H4** groups. Each alternative tends to overvalue or undervalue the expenditures of particular **H** groups and to generate outliers of a specific type. Thus, in the context of the regression model, we can compare the **H** effects and the pattern of outliers after controlling for household characteristics. In particular, outliers attributable in each case to a faulty imputation of reference periods are selected and individually corrected. The three improved versions are compared, and option (c) turns out to be favored again.

The estimation of annual food expenditures from survey data is important in many applications, but here we only focus on the implications of different estimation procedures in two areas. In the first place, like statistical bureaus in other countries, the INE collects the EPF at regular time intervals to estimate the base weights of the official consumer price indexes. 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, biased estimates at the individual level might affect the measurement of household inequality when individual welfare is approximated by total household expenditure.

What are the implications for inflation and inequality measurement of maintaining INE's alternative (a) rather than choosing our preferred option (c)? The main conclusions are the following:

1. Official price indexes can be seen as weighted averages of commodity price changes, with weights equal to average budget shares for those commodities. The differences in the average food share and in the share of food expenditures devoted to specific food items under the two alternatives have a small impact on the measurement of either general or food price inflation.

2. There is a significant reduction in household food-expenditure inequality, ranging from 12% to 50%. For the distribution of household total expenditure, the inequality improvement is maintained but amounts only to 1.5%–3.0%. The results' range of variation depends on alternative decisions about two standard methodological problems in income distribution theory—how to compare food or total expenditure for households of different size and which inequality index should be used.

The rest of the article is organized in four sections and two appendixes. Section 1 presents the data, the notation, the Poisson model for the frequency of bulk purchases, and the three alternatives. Section 2 is devoted to the regression analysis of all alternatives, before and after the correction for outliers directly attributable to their known shortcomings. Section 3 discusses the consequences for inflation and inequality measurement of adopting our preferred alternative *versus* the one originally suggested by the INE. Section 4 contains some concluding remarks. Appendix A is devoted to the description of household characteristics and the regression results for the full model. Appendix B describes a procedure to allocate the aggregate food expenditure among a set of 25 food items for those households who, having made some bulk purchases, did not provide any commodity breakdown—a necessary step prior to the estimation of food inflation rates.

1. DATA, NOTATION, AND THE THREE ALTERNATIVES

1.1 The Available Information on Bulk Purchases

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. Household-budget surveys in Spain are usually rather large. The version for 1990–1991 has 21,155 observations for a population of about 11 million households. We dropped 88 households who were either not observed to make any purchase or were paid in kind. The remaining households are classified into four groups as shown in Ta-

ble 1. The sample and population frequencies, in which the latter are estimated using the blowing-up factors provided by the INE, are given in Table 2.

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**, whereas the remaining 79 belong to **H20**. Similarly, out of the 388 households in **H4**, 321 belong to **H44** and 67 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. Notice that the average weekly expenditure of the sample-week bulk purchase is $m(BP) = 2{,}566$, whereas for the previous three weeks it is $m(PBP) = 4{,}851$.

1.2 The Poisson Model for the Frequency of Purchase

We do not have information about the household distribution into the **F** (frequent), **I** (infrequent), and **N** (never) classes defined in the introduction. 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 **I** 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 $\lambda_1 (> 1)$ and $\lambda_2 (< 1)$. We will call $\nu = \alpha_1\lambda_1 + \alpha_2\lambda_2$ the expected number of purchases in a four-week period according to this model.

Given the available information about the vector of unknown parameters $\theta = (\alpha_1, \alpha_2, \lambda_1, \lambda_2)$, we use the method of moments. John (1970) showed that this method provides an asymptotic normal distribution for the estimators of θ in Poisson mixture models such as the one considered in this article. He also derived the asymptotic covariance matrix of the moment estimators. In this case, we know from Table 2 that

1. the proportion of people who did not make bulk purchases in the four-week period is .7284, so we can write

$$\alpha_1 e^{-\lambda_1} + \alpha_2 e^{-\lambda_2} + (1 - \alpha_1 - \alpha_2) = .7284; \quad (1)$$

2. the proportion of people who did not make bulk purchases in the sample week is

$$\alpha_1 e^{-\lambda_1/4} + \alpha_2 e^{-\lambda_2/4} + (1 - \alpha_1 - \alpha_2) = .9656; \quad (2)$$

Table 1. Household Classification

Variable	Definition	Interpretation
H1	**BP** = **PBP** = 0	No bulk purchases observed
H2	**BP** > 0, **PBP** = 0	Bulk purchases only during the sample week
H3	**BP** = 0, **PBP** > 0	Bulk purchases only during the previous 3 weeks
H4	**BP** > 0, **PBP** > 0	Bulk purchases on both occasions

Table 2. Frequency Distributions by Household Type

Household type	Sample distribution		Population distribution	
H1	15,427	72.2	8,203,138	72.9
H2	404	1.9	193,209	1.7
H3	4,848	23.0	2,670,766	23.7
H4	388	1.9	194,249	1.7
All	21,067	100.0	11,261,362	100.0

3. the proportion of people who did not make bulk purchases in the three weeks before the sample period is

$$\alpha_1 e^{-3\lambda_1/4} + \alpha_2 e^{-3\lambda_2/4} + (1 - \alpha_1 - \alpha_2) = .7456; \quad (3)$$

4. the proportion of people who made one bulk purchase in the sample period is

$$\alpha_1 e^{-\lambda_1/4}(\lambda_1/4) + \alpha_2 e^{-\lambda_2/4}(\lambda_2/4) = .0344. \quad (4)$$

We solve the system of Equations (1)–(4) by a nonlinear optimization routine. An approximate solution (in the least squares sense) to these equations is $\hat{\alpha}_1 = .0353$, $\hat{\lambda}_1 = 1.7678$, $\hat{\alpha}_2 = .4078$, $\hat{\lambda}_2 = .6121$. According to it, **F** households represent 3.5% of the population with an average time between bulk purchases of $4/1.7678 = 2.26$ weeks. For **I** households (roughly 40% of the population), the average time between bulk purchases is $4/.612120 = 6.53$ weeks. The estimated expected number of bulk purchases in the four-week period is given by $\hat{\nu} = \hat{\alpha}_1\hat{\lambda}_1 + \hat{\alpha}_2\hat{\lambda}_2 = .312$, which implies an average time between bulk purchases of $4/.312 = 12.82$ weeks for the population as a whole. 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 $2 \times .0173 + 1 \times .254 + 0 \times .726 = .288$.

The preceding 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 the parameter ν in the range .29 to .36. Solutions differ in the assignment of households to the two classes **F** and **I**, with the corresponding adjustment in the λ parameters. If, for example, $\hat{\alpha}_1$ increases, then $\hat{\lambda}_1$ decreases so that the product is approximately maintained. The particular solu-

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	—	4,769	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

tion already analyzed seems plausible to us and will be used in the sequel.

Although we can compute the covariance matrix of the estimates following John (1970), what we need is the standard error of the estimated mean number of bulk purchases $\hat{\nu}$. Because $\hat{\nu}$ is a nonlinear vector of θ, it can be approximated by Taylor expansion. In this article, however, we use a different approach. Because the range of solutions that can be obtained yields a value for $\hat{\nu}$ in the range (.29, .36), we will carry out a sensitivity analysis of our solution for this range of parameter values.

Without frequency data, we must assume that the distribution of the number of purchases is independent of the amount spent. Then, taking into account that there are 13 periods of four weeks in a year consisting of 52 weeks, the average amount spent on bulk purchases on a yearly basis is equal to 13 $\nu\mu(\mathbf{BP})$, where ν is estimated by $\hat{\nu}$ and $\mu(\mathbf{BP})$—the average bulk-purchase expenditure—is estimated from the available sample data, $m(\mathbf{BP}) = 2,566$ (see Table 3). Therefore, the average amount that must be added to each household on a yearly basis is $13 \times .312 \times m(\mathbf{BP}) = 4.056m(\mathbf{BP})$. For individual groups, the estimated Poisson model implies that we must add $13 \times 1.7678 \times m(\mathbf{BP}) = 22.98m(\mathbf{BP})$ to 3.53% of **F** households and $13 \times .6121 \times m(\mathbf{BP}) = 7.96m(\mathbf{BP})$ to 40.78% of **I** households.

1.3 The Three Alternatives

The three alternatives assign different reference periods to **BP** and **PBP** on the basis of the **H** group to which each household belongs. We do not know the relationship between the **H** groups and the estimated Poisson distribution into 3.53% of **F** households, 40.78% of **I** households, and 55.69% of **N** households. We can safely assume, however, that all **H4** households, representing 1.72% of the population, are in group **F**. The remaining 1.80% of **F** households can be assumed to belong to groups **H2** or **H3**. The rest of groups **H2** and **H3**, 25.43%, can be assumed to be **I** households. This means that at least 16% of **H1** households do acquire bulk purchases occasionally. Because they are not observed to make them at all, their food expenditures are necessarily undervalued in all of the following imputation procedures.

Alternatives (a), (b), and (c) differ in the frequency with which **H2**, **H3**, and **H4** households are assumed to make bulk purchases. The implications about the average additions to be made are reported in Table 4.

Under alternative (a), used by the INE, information on **PBP** is ignored, but a weekly reference period is assigned to **BP**. Apparently, the INE is 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 is offset by the overvaluation of **BP** for **H2** and **H4** households. With this procedure, the INE is adding an average of $52 \times .034 \times m(\mathbf{BP}) = 1.768m(\mathbf{BP})$ so that (1) it is missing more than half of the food-expenditure increment attributable to bulk purchases and (2) it greatly overestimates the increment for a small part of the population.

Under alternative (b), all bulk purchases are assigned a four-week reference period. This means that this procedure adds $13m(\mathbf{BP})$ to **H2** households. From Tables 2 and 3 we obtain that $m(\mathbf{PBP}) = 1.876m(\mathbf{BP})$; therefore, this procedure is adding $24.39m(\mathbf{BP})$ to the 23.72% of the population in group **H3** and $13 \times 2.876m(\mathbf{BP}) = 37.39m(\mathbf{BP})$ to 1.73% of the population in group **H4**. This suggests that groups **H3** and **H4** are probably overvalued. Globally, we are adding on average an additional food expenditure of

$$[.0171m(\mathbf{BP}) + .2372m(\mathbf{PBP}) + .0173(m(\mathbf{BP}) + m(\mathbf{PBP}))]13 = 6.65m(\mathbf{BP}). \quad (5)$$

This results in an overestimation of total expenditure by roughly 50%.

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

$$[.0171m(\mathbf{BP}) + .2372m(\mathbf{PBP}) + .0173(m(\mathbf{BP}) + m(\mathbf{PBP}))]\mathbf{y} = 4.056m(\mathbf{BP}).$$

Taking into account that $m(\mathbf{PBP}) = 1.876m(\mathbf{BP})$, we find that $\mathbf{y} = 7.924$ instead of 13. This implies an average time between bulk purchases equal to $52/7.924 = 6.56$ weeks. In this case, we are adding an average amount of $22.79m(\mathbf{BP})$ to 1.73% of frequent households in **H44**, $7.924m(\mathbf{BP})$ to a small group of infrequent households in **H2**, representing 1.71% of the population, and $14.86m(\mathbf{BP})$ to **H3** households, which constitute 23.72% of the population.

For comparison purposes, in Table 5 we present average weekly expenditures, weekly expenditures per capita, and the share of total expenditures devoted to food by each group and the population as a whole under the three options. It can be seen that alternative (c) produces the smallest variability among the groups. Because weekly food expenditures are not expected to vary much among groups, Table 5 suggests that this alternative is to be preferred. This

Table 4. Addition to Food Expenditures From Bulk Purchases Under Different Alternatives

Alternatives	Average addition to each group: H2 (1.71%)	H3 (23.72%)	H4 (1.73%)	Global average addition
a	$52m(\mathbf{BP})$	—	$52m(\mathbf{BP})$	$1.77m(\mathbf{BP})$
b	$13m(\mathbf{BP})$	$24.39m(\mathbf{BP})$	$37.39m(\mathbf{BP})$	$6.65m(\mathbf{BP})$
c	$7.92m(\mathbf{BP})$	$14.16m(\mathbf{BP})$	$22.79m(\mathbf{BP})$	$4.06m(\mathbf{BP})$
Poisson estimates:	$7.96m(\mathbf{BP})$ to 40.78% of **I** households		$22.98m(\mathbf{BP})$ to 3.53% of **F** households	$4.06m(\mathbf{BP})$

Table 5. Average Weekly Food Expenditures and Mean Food Share

	Weekly expenditures			Weekly expenditure per capita			Food share		
Group	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
H1	11,431	11,431	11,431	3,770	3,770	3,770	.314	.314	.314
H20	28,427	16,507	14,957	7,701	4,380	3,949	.440	.308	.284
H22	16,802	10,878	10,107	4,832	3,103	2,878	.329	.244	.230
H3	12,503	17,312	15,435	3,572	4,991	4,437	.253	.19	.296
H40	29,034	20,698	16,512	9,699	6,146	4,888	.85	.317	.274
H44	17,835	16,017	13,039	5,074	4,530	3,670	.307	.286	.248
All	11,963	12,930	12,414	3,785	4,070	3,919	.300	.314	.307

analysis does not take into account other household characteristics, however, and therefore can be very misleading. In Section 2 we will compare the group means once household differences have been taken into account by regression analysis.

2. REGRESSION ANALYSIS

2.1 First Set of Results for the Three Alternatives

Our first task is to place the previous discussion in a multiple regression setting. Following Deaton, Ruiz-Castillo, and Thomas (1989), we select a flexible functional form for the food-share equation. Taking alternative (a) as the reference option, we have

$$\mathbf{SHa} \equiv \mathbf{Fa}/\mathbf{TEa} = \alpha + \beta \ln(\mathbf{PCTEa}) + \lambda \ln(\mathbf{HS}) + \Sigma_j \delta_j \mathbf{Nj} + \gamma \mathbf{z} + \varepsilon, \quad (6)$$

where **Fa** and **TEa** are household food expenditure and total expenditure, respectively, so that **SHa** is the food share under alternative (a); **HS** is household size, $\mathbf{PCTEa} \equiv \mathbf{TEa}/\mathbf{HS}$ is per capita household total expenditure; $\mathbf{Nj} \equiv \mathbf{HSj}/\mathbf{HS}$, and **HSj** is the number of household members in the jth age bracket; and **z** is a vector of explanatory variables that 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 a set of dummy variables **Hi**, where $i = 20, 22, 3, 40$, and 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 6 we present the coefficient estimates for the variables we are more interested in (with t values in parentheses), total expenditure elasticities, and a measure of the goodness of fit. As a measure of the heterogeneity of the **H** groups among the three alternatives, we have included the Euclidean distance from 0 of the estimated **H** coefficients.

The following comments are in order:

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

2. For the sample as a whole, food is clearly a necessity, with a total expenditure elasticity of approximately .65 under all options.

3. Option (c) seems to be the one that produces more homogeneity among the **H** groups. The $\Sigma \mathbf{H}_i^2$ is half of that of option (b) and almost one-tenth that of option (a).

4. As expected, **H3** households appear undervalued in option (a), which does not give any weight to **PBP**. On the contrary, because **BP** are treated as weekly expenditures, groups 20, 22, 40, and 44 appear very significantly overvalued. Households in **H20** and **H40**, who could not remember their allocation of bulk purchases to specific commodities, seem to exaggerate the amount spent on food, a fact already apparent in Table 3 when we compare their average expenditure on **BP** to that of groups **H22** and **H44**, respectively. This might mean that forgetful households tend to think that they spent more on bulk purchases than households who keep good records of it. On the other hand, although group **H40** has a larger average **BP** than group **H44**, the two groups' **PBP** values are rather close to each other. This might be the case because **H44** households tend to suffer also from an idealization of the past effect. Consequently,

Table 6. Summary of Regression Results for Different Options

Selected variables	Option a	Option b	Option c
INTERCEPT	1.7191 (75.1)	1.8269 (79.5)	1.7999 (79.2)
H3	−.0201 (−10.9)	.0580 (31.4)	.0298 (16.3)
H20	.1664 (13.6)	.0121 (1.0)	−.0158 (−1.3)
H22	.0524 (8.1)	−.0453 (−7.1)	−.0614 (−9.6)
H40	.1718 (12.4)	.0969 (7.1)	.0454 (3.3)
H44	.0505 (8.1)	.0284 (4.6)	−.0164 (−2.6)
ln **PCTE**	−.1022 (−61.9)	−.1097 (−66.3)	−.1079 (−65.8)
Elasticity	.6597	.6504	.6487
R^2	.4054	.4027	.4041
Sample size	21.063	21.067	21.067
$\Sigma \mathbf{H}_i^2 10^4$	629	152	72

there is no surprise in the fact that groups **H20**, **H40**, and **H44** appear as particularly overvalued under option (a).

5. Notice that the vast majority of **H3** households are possibly 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** households. Table 3 indicates that the group **H3** is much closer on average to group **H20**. Therefore, we might conjecture that, just as we saw with groups **H20**, **H40**, and **H44**, because of a certain idealization of the past the bulk purchases in group **H3** are also exaggerated. This, together with the fact that option (b) assumes a short average period between bulk purchases, explains why **H3** appears overvalued under this option in Table 6. The amount of overvaluation, however, is one-third that of **H20** and **H40** under option (a). On the other hand, group **H22** is now significantly undervalued. Taking into account Table 3, we conjecture that these infrequent bulk purchasers 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 as much as under alternative (a). Finally, note that, as expected, the intercept is larger in (b) than in (a) because where there was overall underestimation we have now overestimation relative to the prediction of the Poisson model.

6. 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 essentially insignificant, but with a minus sign, whereas the **H22** group appears heavily undervalued. As expected, the intercept in this option is between that of options (a) and (b). Possibly the best feature of option (c) with respect to option (b) is that the large group of **H3** households is now much less overvalued.

2.2 Correction for Outliers

We have seen how a priori views concerning undervaluation and overvaluation in each of the three alternatives were confirmed by the regression analysis. Therefore, we have grounds to select those outliers that can be attributed to imperfect imputation of bulk purchases to correct them on an individual basis to reach a second, presumably improved version of each alternative.

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 a 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 removed 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 undervalued; that is, the observed response value at these n_0 points is $y_{\text{ob}} = y_{\text{real}} - k_i$, 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 signs opposite to that of the dummy variable in the group (see Table 6 for the latter). Because group **H1** may be undervalued 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 of 5 for the studentized residual. This high value has been chosen for the following reasons: (1) correction for small effects is to be avoided because, as explained before, the bias of the intercept may lead to a biased estimation; (2) outliers due to a wrong imputation for bulk purchases are expected to be large; and (3) 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 7 (outliers for option (b) are available on request). The correction of these outliers leads to what we call versions (aa), (bb), and (cc). The results are summarized in Table 8, and the full model for version (cc), very similar to the other versions, appears as Model 2 in Appendix A.

The main implications of these corrections are as follows:

1. The coefficient of the log of household size is the only one that changes notably, becoming significant under the three options. As expected, goodness of fit is substantially improved, with an R^2 of approximately .46 for all alternatives, up from .40 before outlier corrections. Moreover, the t values are generally improved.

2. Total expenditure elasticity for the full sample goes down, approximately, from .65 to .62 in all alternatives.

3. The largest reduction of heterogeneity appears in option (aa), in which the Euclidean distance from 0 of the **H** variables is now half that in option (a). The most homogeneous option, however, is again (cc) with a heterogeneity statistic half that of option (bb).

4. In option (aa), even after adjusting for outliers **H3** households still appear significantly undervalued, but all

Table 7. 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

Table 8. Summary of Regression Results Under Different Options: The Full Sample

Selected variables	*Option (aa)*	*Option (bb)*	*Option (cc)*
INTERCEPT	1.9028 (87.4)	1.9789 (91.9)	1.9697 (91.7)
H3	−.0165 (−9.5)	.0491 (28.6)	.0241 (14.1)
H20	.1241 (10.8)	.0111 (1.0)	−.0160 (−1.4)
H22	.0408 (6.7)	−.0484 (−8.1)	−.0616 (−10.3)
H40	.1140 (8.7)	.0900 (7.0)	.0368 (2.9)
H44	.0441 (7.5)	.0247 (4.2)	−.0192 (−3.2)
ln **PCE**	−.1149 (−73.2)	−.1196 (−77.1)	−.1192 (−77.0)
Elasticity	.6231	.6234	.6178
R^2	.4604	.4582	.4631
Sample size	21.039	21.040	21.039
$\Sigma \mathbf{H}_i^2 10^4$	323	136	64

the rest, especially groups **H20** and **H40**, remain seriously overvalued.

5. 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**.

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

3. IMPLICATIONS

Having done 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).

3.1 The Measurement of Inflation

We have measured the inflation for the food category during 1993 and 1994 under both alternatives. For that purpose, as in the official system, we have constructed a Laspeyres-type price index for the population as a whole (including those households that did not enter into the regression analysis). Let Fa^h be the food expenditure of household h under alternative (a) for example, and let w_i^h be the share of Fa^h (net of unclassifiable expenditures) devoted to food item $i = 1, \ldots, 25$. Let $W = (W_1, \ldots, W_{25})$ be the 25-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 Fa^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 a vector under alternative (cc) is described in Appendix B.

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 .2996 and .3108 for options (a) and (cc), respectively. This is not a large difference either. Therefore, we should not expect large 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, whereas under alternative (cc) they are 105.24 and 110.22 for those same years.

3.2 The Measurement of Inequality

Households with different characteristics have different needs, so their incomes or expenditures are not directly comparable. In this article we select household size, s^h, as the characteristic most likely to create differences in needs. To compare the food expenditures of households with different sizes under alternative (a), for example, define adjusted food expenditure by

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

This is a convenient parameterization, which covers the range from the extreme case in which no adjustment is made for household size, $\Theta = 0$, to the case in which what is assumed to be comparable across household sizes is per capita household food expenditure when $\Theta = 1$. In general, we expect to find some economies to scale in consumption within the household. Therefore, we also study the intermediate case $\Theta = .5$.

Because of its good properties, we have considered the generalized entropy family of relative inequality indexes [For a characterization, see Shorrocks (1980). For a defense, discussion, and applications, see Cowell (1984), Coulter, Cowell, and Jenkins (1992a,b), and Ruiz-Castillo (1995)]. This family is defined by

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

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

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

where $\mu(\mathbf{z})$ is 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 square of the coefficient of variation—and a member more sensitive to the lower part, $c = -1$. We have also estimated the two indexes originally suggested by Theil corresponding to $c = 1$ and $c = 0$.

The results are in the left side of Table 9. 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% to a maximum of 50%. Such an improvement is greater at an intermediate value of the parameter Θ and also greater the more sensitive one is to the upper tail of the distribution.

Finally, we have carried on the same exercise for the distribution of total expenditure. The results are in the right side of Table 9. The improvement in inequality persists in

Table 9. Inequality Under Different Options

Options	Food expenditure inequality				Total expenditure inequality			
	$c = 2$	$c = 1$	$c = 0$	$c = -1$	$c = 2$	$c = 1$	$c = 0$	$c = -1$
				$\Theta = .0$				
Option (a)	.1813	.1636	.1853	.3163	.2525	.2046	.2169	.3089
Option (cc)	.1613	.1463	.1593	.2185	.2474	.2021	.2134	.2994
(a)/(cc)	1.1240	1.1182	1.1632	1.4476	1.0206	1.0123	1.0164	1.0317
				$\Theta = .5$				
Option (a)	.1412	.1249	.1341	.1982	.2128	.1701	.1697	.2111
Option (cc)	.1208	.1066	.1089	.1308	.2094	.1674	.1664	.2043
(a)/(cc)	1.1689	1.1717	1.2314	1.5145	1.0162	1.0161	1.0198	1.0333
				$\Theta = 1.0$				
Option (a)	.1726	.1414	.1423	.1887	.2575	.1922	.1831	.2179
Option (cc)	.1497	.1224	.1184	.1349	.2535	.1894	.1800	.2123
(a)/(cc)	1.1530	1.1552	1.2018	1.3988	1.0158	1.0148	1.0172	1.0264

this domain but loses importance: The range of variation is from 1.5% to 3.0%.

The implications for inflation and inequality just reviewed have been obtained for a version of option (c) in which the average time between bulk purchases is 6.56 weeks. As we saw in Section 1, this is the period that results from adding an average expenditure on account of bulk purchases equal to the amount implied by the Poisson model in the case in which the expected number of bulk purchases in the four-week observation period is estimated to be $\nu = .312$. This leads to a rather long estimated average time between bulk purchases of 12.82 weeks. Therefore, in our sensitivity analysis we have considered the upper bound for $\nu = .36$, which implies an average time between bulk purchases of 5.68 weeks in a new option (c).

The implications of working under this upper bound are essentially the same as before: (1) Changing from option (a) to (cc) has little effect on the measured general inflation rate or the rate for food. (2) There is a considerable effect on the measurement of food inequality and a much smaller effect on the measurement of total expenditure inequality. The reduction in inequality is only slightly greater than the reduction reported previously under option (cc).

4. CONCLUDING REMARKS

The increasing popularity of bulk purchases among certain strata of the population makes it more difficult to collect information on food purchases and to estimate the annual food expenditures of each household. In our treatment of this problem with data from the Spanish 1990–1991 EPF, we have demonstrated that it is always preferable to use as much of the available information as possible, however incomplete it might be. Moreover, we have shown how to improve our estimation of annual food expenditures by modeling a crucial parameter for which there is no direct knowledge—the frequency of bulk purchases.

A simple Poisson model for the frequency of bulk purchases by different population subgroups allows us to ascertain the extent to which different imputation strategies overvalue or undervalue the average food expenditure due to bulk purchases—option (a), currently used by the INE, which ignores much of the available information; option (b), which uses all the information in a rough way; and option (c) which, in addition, makes good use of the implications of the Poisson model.

The three alternatives give rise to predictable outlier patterns in a regression model of the food share as a function of total expenditure and household characteristics. We have shown how the correction for individual outliers improves all options in the sense of reducing the heterogeneity across the subgroups. Moreover, both before and after the correction for outliers, we have seen that option (c) improves the treatment of most groups and reduces the amount of heterogeneity among them.

For the construction of a food price index to measure the rate of inflation, we need to estimate the average share of food expenditure devoted to a set of 25 food items. A number of households do not provide information on how they allocate their bulk purchases among the food items, some because they were not questioned by the INE and some because they could not recall such detail.

In Appendix B we show how to use the available information to solve this allocation problem too. First, we find a reasonable partition of the commodity space into what we call "bulk-purchase goods," "weekly goods," and "other goods." Goods are classified according to their prominence, respectively, within the bulk purchases, within the weekly smaller acquisitions, or in neither in the budget of those households for which we have complete information. Then, by means of regression analysis we confirm that, relative to these commodity subsets, all household groups behave in general agreement with our expectations based on evidence from their aggregate food behavior. Finally, we justify our way of partitioning bulk purchases into specific food categories because it tends to raise the affected households' imputed share of bulk-purchase goods and to lower their share of weekly goods.

The full exploitation of all available information with the help of a Poisson model and regression analysis, as attempted in this article, is more important for some purposes than for others. When we study some implications of adopt-

ing INE's option (a) versus our preferred alternative (c) we find little difference as far as the measurement of inflation is concerned but a considerable difference in food and total expenditure inequality. This result is robust to different specifications of option (c), depending on different estimates of the parameters in the Poisson model.

ACKNOWLEDGMENTS

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 assistance. Financial aid from the Fundación Caja de Madrid, from Projects PB93-0230 and PB96-0111 of the Spanish DGICYT, and from the Cátedra BBV de Calidad is gratefully acknowledged.

APPENDIX A: VARIABLES AND REGRESSION RESULTS

A.1 Definitions of Variables

Demographic

HS = household size

Nj = $\mathbf{HS}_j/\mathbf{HS}$, 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 spouse
SED1* = spouse educational level: illiterate, without formal studies, first and second grade
SED2 = high school

Table A.1. Means of Selected Continuous Variables

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

SED3 = college degree and graduate studies

SOCIO1 = agrarian working class and small landowners
SOCIO2* = nonagricultural working class and other unclassifiable members of the labor force
SOCIO3 = agrarian entrepreneurs, armed forces, nonagrarian entrepreneurs 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

Table A.2. 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	.3	.6	.6	—	.4
	100.0	100.0	100.0	100.0	100.0
NEARN					
0	.06	—	—	—	.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

Table A.3. Model 1: Dependent Variable: Food Share Under Alternative (c)

INTERCEPT	1.7999 (79.2)	**SQM**	−.0001 (−6.9)
H3	.0298 (16.3)	**TEN2**	.0343 (11.1)
H20	−.0158 (−1.3)	**TEN3**	.0445 (11.4)
H22	−.0614 (−9.6)	**TEN4**	.0417 (11.8)
H40	.0454 (3.3)	**TEN5**	.0093 (3.1)
H44	−.0164 (−2.6)	**BUILD2**	−.0133 (−3.6)
ln **PCTE**	−.1079 (−65.8)	**BUILD3**	−.0136 (−6.4)
ln **HS**	−.0026 (−.8)	**NSRY**	−.0284 (−12.2)
N1	−.0327 (−3.1)	**MUN1**	.0269 (7.4)
N2	−.0509 (−5.6)	**MUN2**	.0159 (5.0)
N3	−.0342 (−4.2)	**MUN3**	.0105 (3.7)
N4	−.0719 (−7.4)	**MUN4**	.0091 (3.3)
N5	−.0797 (−12.3)	**MUN5**	.0125 (4.9)
N6	−.0466 (−9.6)	**MUN6**	.0076 (2.8)
N7	.0071 (1.7)	**CCAA3**	−.0096 (−2.2)
N8	.0132 (3.3)	**CCAA4**	−.0096 (−7.2)
NEARN	−.0057 (−7.4)	**CCAA7**	−.0396 (−2.0)
S	.0089 (3.0)	**CCAA8**	−.0064 (−4.4)
HHED1	.0142 (3.8)	**CCAA10**	−.0164 (−6.4)
HHED3	−.0041 (−1.8)	**CCAA11**	−.0358 (−7.9)
HHED4	−.0156 (−5.5)	**CCAA12**	.0306 (10.0)
HHED5	−.0196 (−4.8)	**CCAA13**	−.0193 (−7.8)
HHED6	−.0245 (−5.3)	**CCAA15**	−.0269 (−4.1)
SED0	−.0238 (−7.6)	**CCAA16**	−.0107 (−3.1)
SED2	−.0060 (−1.8)	**WINTER**	−.0070 (−3.3)
SED3	−.0077 (−4.2)	**SUMMER**	.0041 (2.0)
SOCIO1	.0109 (3.2)	**AUTUMN**	−.0096 (−4.6)
SOCIO3	−.0062 (−2.8)	**WEEK2**	−.0026 (−1.7)
SOCIO4	−.0077 (−2.4)	**WEEK3**	.0075 (2.7)
SOCIO5	.0141 (5.5)		
MIGR	.0083 (2.3)		
R^2	.4041		
Sample size	21.067		

NOTE: All variables with at least a 1.70 *t* value in absolute terms in Model 1 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 also has a significant negative effect. For the household head, the greater the educational level attained, the smaller the food share. The effect of the spouse's educational level, whenever present, is less clear. Lower socioeconomic classes and recent inmigrants have significantly higher 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 cities Ceuta and Melilla are insignificantly different from the mean. The quarter and/or the week in which the survey took place has no clearly interpretable effect.

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 = nonresidential 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

Table A.4. Model 2: Dependent Variable: Food Share Under Alternative (cc)

INTERCEPT	1.9697 (91.7)	**SQM**	−.0001 (−7.7)
H3	.0241 (14.1)	**TEN2**	.0348 (12.1)
H20	−.0160 (−1.4)	**TEN3**	.0427 (11.7)
H22	−.0614 (−10.3)	**TEN4**	.0410 (12.4)
H40	.0368 (2.9)	**TEN5**	.0109 (4.0)
H44	−.0192 (−3.3)	**BUILD2**	−.0151 (−4.4)
ln **PCTE**	−.1192 (−77.0)	**BUILD3**	−.0140 (−7.0)
ln **HS**	−.0145 (−4.6)	**NSRY**	−.0265 (−12.2)
N1	−.0306 (−3.1)	**MUN1**	.0148 (5.0)
N3	−.0302 (−3.9)	**MUN3**	.0119 (4.4)
N4	−.0706 (−7.8)	**MUN4**	.0083 (3.2)
N5	−.0755 (−12.5)	**MUN5**	.0122 (5.1)
N6	−.0521 (−11.5)	**MUN6**	.0102 (3.9)
N7	.0047 (1.2)	**CCAA3**	−.0099 (−2.4)
N8	.0109 (2.9)	**CCAA4**	−.0376 (−7.3)
NEARN	−.0049 (−4.9)	**CCAA7**	−.0073 (−2.5)
S	.0028 (1.0)	**CCAA8**	−.0185 (−5.3)
HHED1	.0165 (4.7)	**CCAA10**	−.0178 (−7.3)
HHED3	−.0043 (−2.0)	**CCAA11**	−.0432 (−10.3)
HHED4	−.0134 (−5.1)	**CCAA12**	.0337 (11.8)
HHED5	−.0201 (−5.3)	**CCAA13**	−.0187 (−8.1)
HHED6	−.0243 (−5.7)	**CCAA15**	−.0265 (−4.3)
SED0	−.0175 (−6.0)	**CCAA16**	−.0102 (−3.2)
SED2	−.0020 (−0.6)	**WINTER**	−.0070 (−3.6)
SED3	−.0140 (−3.5)	**SUMMER**	.0042 (2.2)
SOCIO1	.0103 (3.2)	**AUTUMN**	−.0096 (−4.6)
SOCIO3	−.0058 (−2.8)	**WEEK2**	−.0026 (−1.8)
SOCIO4	−.0047 (−1.5)	**WEEK3**	.0062 (2.4)
SOCIO5	.0131 (5.5)		
MIGR	.0081 (2.4)		
R^2	.4631		
Sample size	21.039		

NOTE: The most important difference is in the coefficient of the log of household size, ln **HS**, which is now clearly significant though 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.

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 *.

A.2 Descriptive Statistics

The means of selected continuous variables are shown in Table A.1. The percentage distributions of selected discrete variables are shown in Table A.2.

A.3. Regression Results

Model 1 is shown in Table A.3 and Model 2 is shown in Table A.4.

APPENDIX B: THE ALLOCATION OF FOOD EXPENDITURE AMONG SPECIFIC ITEMS FOR HOUSEHOLDS WHO DO NOT PROVIDE THAT DETAIL

Option (cc) provides the best possible estimation of annual food expenditures using all the available information. For **H20** and **H40** households, however, bulk purchases made during the sample week must be allocated among the 25 specific food items. The same must be done for bulk purchases during the prior three weeks for **H3** 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 or in a shopping mall, as in smaller acquisitions during weekly errands in the surrounding neighborhood. 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.

Table B.1. 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	.709	35.4	−.0055	−.0195	*	*	*	.0507
2. Prep. fish	1.010	37.8	*	*	.0093	−.0177	*	.0450
3. Prep. vegts.	.672	19.7	−.0025	*	*	*	*	.0203
4. Other foods	.916	27.1	*	*	*	*	*	.0353
5. Coffee, tea, cocoa, etc.	.729	13.9	−.0023	*	*	*	*	.0390
6. Other meats	.750	92.7	*	−.0245	*	*	*	.0643
7. Milk prods.	.718	43.2	−.0025	*	*	−.0199	*	.0336
8. Sugar	.366	6.6	−.0018	−.0039	−.0022	−.0045	*	.0587
9. Fruit preserves	.786	9.5	*	*	*	*	*	.0171
			Weekly					
10. Bread	.096	65.2	.0037	*	*	.0185	*	.3284
11. Fresh vegts.	.568	45.1	.0020	*	*	*	*	.0883
12. Potatoes	.439	18.0	*	*	*	*	*	.0972
13. Fresh fruit	.575	81.3	*	*	−.0119	.0628	*	.1023
14. Eggs	.343	18.8	*	*	*	*	*	.0413
15. Fresh and frozen fish	.752	69.2	*	*	*	*	*	.0874
16. Unclassifiable	1.406	24.4	*	.0526	*	*	*	.0420
17. Grains	.780	57.3	*	*	*	*	*	.0441
			Other					
18. Beef	.846	62.1	*	.0297	*	*	*	.1500
19. Lamb	.932	22.5	*	*	*	*	*	.0729
20. Pork	.562	31.5	*	*	*	*	*	.0744
21. Chicken	.394	43.1	*	*	*	*	*	.0411
22. Milk	.344	68.3	−.0052	−.0246	−.0094	*	−.0093	.1065
23. Non-alc. drinks	.820	19.6	*	*	*	*	*	.0510
24. Alcoholic drinks	.980	31.2	*	*	.0111	*	.0088	.0423
25. Tobacco	.593	56.4	.0064	.0286	*	.0415	*	.1547

For every $i = 1, \ldots, 25$, 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 nine bulk-purchase goods, eight weekly goods, and eight other goods. This is a reasonable classification: (1) Prepared goods of all sorts appear prominently in bulk purchases; (2) all types of fresh items appear as weekly goods; (3) meats of different types, milk, alcoholic and nonalcoholic drinks, as well as tobacco, which is only bought in special stores, are among the other goods.

In the next step, before deciding on an allocation procedure for the preceding 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 only have information on **SE** expenditures. Nevertheless, we run two types of regressions for the sample of 21.039 observations that remain after the outlier analysis leading to option (cc). In the first place, we run 25 regressions to compute total expenditure elasticities for each good. These are presented as column (1) in Table B.1. 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 B.1. Regression coefficients for the five groups, relative to the **H1** reference group, are presented in columns (3) to (7). Insignificant coefficients are singled out by means of an asterisk. Finally, each equation's R^2 is provided in column (8).

1. 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 preceding classification, we expect them to be short of bulk-purchase goods, long on weekly goods, and close to the reference group in other goods. Not counting tobacco, **H3** households satisfy the expected pattern in 13 cases, present a single violation in other goods, and show insignificant coefficients in the remaining 10 cases.

2. 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.

3. 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 six goods and from **H40** in five. In 9 out of these 11 cases, they behave as expected.

4. If the behavior of frequent bulk purchasers in **H44** were well captured by the regression model, their dummy variables would be insignificant. 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.

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 among the 25 items according to the population means. Essentially, we correct **H3**, **H20**, and **H40** households in an appropriate direction. Given that they made bulk purchases 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.

[Received October 1995. Revised September 1997.]

REFERENCES

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

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

John, S. (1970), "On Analyzing Mixed Samples," *Journal of the American Statistical Association*, 65, 755–763.

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

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

Ruiz-Castillo, J. (1995), "The Anatomy of Money and Real Inequality in Spain, 1973–74 to 1980–81," *Journal of Income Distribution*, 4, 265–281.

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.