The Stata Journal (2017)
17, Number 1, pp. 1–18

Estimation of unit values in household expenditure surveys without quantity information

Martina Menon
University of Verona
Department of Economics
Verona, Italy
martina.menon@univr.it

Federico Perali
University of Verona
Department of Economics and
Centre for Household, Income, Labour and Demographic Economics
Verona, Italy
federico.perali@univr.it

Nicola Tommasi
University of Verona
Interdepartmental Centre of Economic Documentation
Verona, Italy
nicola.tommasi@univr.it

Abstract. In this article, we present the `pseudounit` command, which estimates pseudounit values in cross-sections of household expenditure surveys without quantity information. Household surveys traditionally record only expenditure information. The lack of information about quantities purchased precludes the possibility of deriving household-specific unit values. We use a theoretical result developed by Lewbel (1989, *Review of Economic Studies* 56: 311–316) to construct pseudounit values by first reproducing cross-sectional price variation, then adding this variability to the aggregate price indexes published by national statistical institutes. We illustrate the method with an example that uses a time series of cross-sections of Italian household budgets.

Keywords: st00!!, pseudounit, unit values, cross-section prices, demand analysis

1 Introduction

This article presents the theory used to implement `pseudounit`, a command that estimates unit values in cross-sections of household expenditure surveys without quantity information, and describes how the command should be used. Empirical works on demand analysis generally rely on the assumption of price invariance across households, supported by the hypothesis that, in cross-sectional data, there are neither time nor

spatial variations in prices. According to this assumption, each family pays the same prices for homogeneous goods. Microdata with this characteristic allow researchers to estimate only Engel curves without accounting for price effects, which are crucial for both behavioral and welfare applications. Slesnick (1998, 150) states that "the absence of price information in the surveys creates special problems for the measurement of social welfare, inequality and poverty. ... Most empirical work links micro data with national price series on different types of goods[,] so cross sectional variation is ignored. Access to more disaggregate information on prices will enhance our ability to measure social welfare, although it remains to be seen whether fundamental conclusions concerning distributional issues will be affected". In empirical works, such limitation is usually bypassed by analyzing time series of cross-sections where price information comes from aggregate time-series data. Plausible estimates of price effects require a sufficiently long series of cross-sections and, if possible, aggregate price indexes that vary by month and location, usually by region or province.

Household budget surveys of both developed and developing countries can be classified into two broad categories in increasing order of frequency of occurrence: 1) surveys of expenditure and quantities purchased and 2) surveys of expenditure data only. In the first case, where quantities and expenditure are both observed, cross-sectional prices are obtained as implicit prices, dividing expenditure by quantities, and are more properly referred to as "unit values". When dealing with these surveys, one should remember that a proper use of unit values in econometric analyses must account for problems arising from the fact that unit values provide useful information about prices, but differ from market prices in many respects. The ratio between expenditure and quantities bought embed information about the choice of quality (Deaton 1987, 1989, 1988, 1990, 1998; Perali 2003). The level of the unit value of a composite good depends on the relative share of high-quality items and the composition of the aggregate good. Unit values can also be highly variable for supposedly homogeneous goods because the market offers many different grades and types.

On the other hand, when one deals with surveys that report only expenditure information, aggregate national price indexes are usually merged with household expenditure to obtain estimates of price elasticities. Unfortunately, this approach requires a long time series of cross-sectional data to estimate a demand system with sufficient price variation and relies on very restrictive assumptions (Frisch 1959), which often turn out to be rejected in empirical applications. Aggregate price indexes are generally highly correlated, may suffer from endogeneity problems (Lecocq and Robin 2015), and the estimated elasticities are often not coherent with the theory (Atella, Menon, and Perali 2004; Coondoo, Majumder, and Ray 2004; Dagsvik and Brubakk 1998; Lahatte et al. 1998). Thus surveys gathering exclusively expenditure data, such as the Italian household budget survey conducted by the National Statistical Institute (Italian Statistical Institute [ISTAT]) used in our example and the majority of existing household budget surveys, have limited applicability in modern demand and welfare analysis. Thus one should devise an appropriate procedure to compute pseudounit values using the information traditionally available in expenditure surveys, such as budget shares and demographic characteristics, which help reproduce the distribution of the unit-value

variability as closely as possible. The theoretical background for this undertaking is provided in a study by Lewbel (1989).

The remainder of the article is organized as follows: Section 2 presents the theory and method used to derive consumer price indexes and pseudounit values (PUVs) when the main objective is to implement demand analysis of household budget data without quantity information. Section 3 provides the syntax and options of the `pseudounit` command. Section 4 illustrates the application. Section 5 concludes.

2 The estimation of unit values in cross-section analysis of household budget surveys

We introduce a method that recovers unit values when only expenditure information is available using knowledge about aggregate price indexes available from national statistics.

First, as illustrated in section 2.1, we need to collect the consumer price indexes available from official statistics and associate them with each household in the survey. Then, to improve the precision of the estimated price elasticities as shown in Atella, Menon, and Perali (2004), we reproduce as best as we can the price variation of actual unit values, which could be obtained as the ratio between expenditure and quantities if quantity information were available in the survey. The estimation of PUVs is described in section 2.2 and 2.3.

2.1 Consumer price indexes

Eurostat adopts the classification of individual consumption by purpose (COICOP), which is a nomenclature developed by the United Nations Statistics Division to classify and analyze individual consumption expenditures incurred by households, nonprofit institutions serving households, and general government according to their purpose.[1] National statistical institutes traditionally publish consumer price indexes per each COICOP category monthly, which are collected at the provincial level.

Let P_{rm}^{ij} be the consumer price index for the jth of the ith COICOP group with $i = 1, \ldots, n$ collected monthly by national statistical institutes, $m = 1, \ldots, M$, per each territorial level $r = 1, \ldots, R$, such as a province or a region. These price indexes are the same for all households living in the same region and interviewed in the same month. If detailed price information is disaggregated by territorial level or time is not available, then we have only P^i, which is the same for all households. With this highly limited price information, demand analysis cannot be implemented, because the data matrix is not invertible and PUVs must be estimated.

1. The COICOP top-level aggregation encompasses 12 categories: food and nonalcoholic beverages; alcoholic beverages, tobacco and narcotics; clothing and footwear; housing, water, electricity, gas and other fuels; furniture; health; transportation; communication; recreation and culture; education; restaurants and hotels; and miscellaneous goods and services.

The next task is to match the monthly price index specific to each territorial unit, P_{rm}^{ij}, with all households living in province or region r and interviewed at month m. Then, each P_{rm}^{ij} is aggregated into a price index P_{rm}^{i} for $i = 1, \ldots, K$ groups corresponding to the goods selected for the empirical demand analysis. The aggregation uses Laspeyres indexes

$$P_{rm}^{i} = \sum_{j=1}^{n_i} \left(P_{rm}^{ij} w_{ij} \right) \tag{1}$$

where $j = 1, \ldots, n_i$, with n_i being the number of goods within group i and w_{ij} being the weights provided by national statistical institutes for each item j of group i.[2] As an example, we may suppose that a budget is divided into $i = 1, 2$ groups such as food and nonfood and that the food subgroup is composed by $j = 1, 2, 3$ items such as cereals, meat, and other food.[3]

So far we have described how to prepare the data matrix containing information about the available price indexes. Next, we present the background theory used to pursue the objective to reconstruct the cross-sectional variability of unit values.

2.2 Demographically varying PUVs: Theory

Lewbel (1989) proposes a method to estimate the cross-sectional variability of actual unit values by exploiting the demographic information included in generalized "within-group" equivalence scales or, more generally, demographic functions.[4] For a group i of goods, these are defined as the ratio of a subutility function of a reference household to the corresponding subutility function of a given household estimated without price variation in place of "between-group" price variation. The method relies on the assumption that the original utility function is homothetically separable and "within-group" subutility functions are Cobb–Douglas.

Consider a separable utility function $U\{u_1(q_1, d), \ldots, u_n(q_n, d)\}$ defined over the consumption of good q_i and a set of demographic characteristics d, where $U(u_1, \ldots, u_n)$ is the "between-group" utility function and $u_i(q_i, d)$ is the "within-group" subutility function, where $i = 1, \ldots, n$ denotes the aggregate commodity groups. Demographic characteristics, d, affect U indirectly through the effects on the within-group subutility function. Define the group equivalence scale, $M_i(q, d)$, as

$$M_i(q_i, d) = \frac{u_i(q_i, d^h)}{u_i(q_i, d)}$$

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

$$U(u_i, \ldots, u_n) = U\left(\frac{Q_1}{M_1}, \ldots, \frac{Q_n}{M_n} \right)$$

2. When not available, the subgroup budget shares can be used as the weight of aggregation.
3. If the interest is to build a time-series collection of cross-sections of household budget surveys, then in the base year, the indexes for all goods are equal to 100.
4. This section closely reproduces the procedure developed by Lewbel (1989).

which is formally analogous to Barten's (1964) technique to introduce demographic factors in the utility function. Define further the price index for group i as $P_i = Y_i^h/Q_i$, where Y_i^h is the expenditure on group i by the reference household. To guarantee group demands are closed under unit scaling, we must apply a scaling factor k_i to the quantity index, Q_i, that makes $P_i = 1$ for all i when $p_{ij} = 1$ for all i and j. This would occur, for example, in a base year when p_{ij} are in index form. Thus $P_i = Y_i^h/k_i Q_i$. Barten's utility structure implies the following share demands for each household with total expenditure Y:

$$W_i = H_i(P_1 M_1, \ldots, P_n M_n, Y)$$

This takes the form of $W_i^h = H_i^h(P_1, \ldots, P_n, Y^h)$ for the reference household with scales $M_i = 1$ for all i. The further assumption of homothetic separability admits two-stage budgeting (Deaton and Muellbauer 1980) and implies the existence of functions V_i, such that $P_i = V_i(p_i, d^h)$ is the price index of group i for the reference household with demographics d^h, $p_i = (p_{i1}, \ldots, p_{in_i})$ is the vector of prices, where n_i is the number of goods that compose group i. By analogy with the definition of group equivalence scales in utility space, it follows that

$$M_i = \frac{V_i(p_i, d)}{V_i(p_i, d^h)}$$

where $V_i(p_i, d) = M_i P_i$. Therefore, when demands are homothetically separable, each group scale depends only on relative prices within group i and on d as expected given that homothetic separability implies strong separability. Maximization of $u_i(q_i, d)$, subject to the expenditure $p_i q_i = Y_i$ of group i, gives the budget share for an individual good $w_{ij} = h_{ij}(p_i, d, Y_i)$. For homothetically separable demands, the budget shares do not depend on expenditure $w_{ij} = h_{ij}(p_i, d)$ and integrate back in a simple fashion to $V_i = M_i P_i$. This information can be used at the between-group level in place of price data to estimate $W_i = H_i(V_1, \ldots, V_n, Y)$. Under the assumption that the subgroup utility functions are Cobb–Douglas with parameters specified as "shifting" functions of demographic variables alone, we can specify the following relationship:

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

The shares $w_{ij} = (\partial \log V_i)/(\partial \log p_{ij})$ then correspond to the demographic functions

$$w_{ij} = h_{ij}(p_i, d) = m_{ij}(d)$$

with

$$\sum_{j=1}^{n_i} w_{ij}(d) = \sum_{j=1}^{n_i} m_{ij}(d) = 1$$

The implied price index is

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

with

$$k_i = \prod_{j=1}^{n_i} m_{ij} \left(d^h\right)^{-m_{ij}(d^h)}$$

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

Note[5] that, assuming separable and homothetic preferences within groups and letting q_{ij} denote scaled units so that corresponding prices p_{ij} are unity in a base year,[6] the group cost function for the reference household is

$$c_i\left(u_i, \widetilde{p}_i, d^h\right) = k_i u_i\left(\widetilde{q}_i, d^h\right) \frac{b\left(\widetilde{p}_i, d^h\right)}{k_i} = k_i Q_i P_i$$

where $b\left(\widetilde{p}_i, d^h\right)$ is concave and linearly homogeneous in prices, and time subscripts are omitted for simplicity. To ensure the group price index is unity in the base year, we denote the scaling factor as $k_i = b\left(1, d^h\right)$. Thus $P_i = \{b\left(\widetilde{p}_i, d^h\right)\}/k_i = V_i\left(\widetilde{p}_i, d^h\right)$, and the price per equivalent capita is

$$M_i P_i = \frac{b\left(\widetilde{p}_i, d^h\right)}{k_i} \frac{b\left(\widetilde{p}_i, d\right)}{b\left(\widetilde{p}_i, d^h\right)} = \frac{b\left(\widetilde{p}_i, d\right)}{k_i} = V_i\left(\widetilde{p}_i, d\right)$$

When the subutility functions are Cobb–Douglas, $b\left(\widetilde{p}_i, d\right) = \prod_{j=1}^{n_i}\left(\widetilde{p}_{ij}/m_{ij}\right)^{m_{ij}}$, and it is easy to see that the scaling factor is $k_i = \prod_{j=1}^{n_i}\left(1/m_{ij}^h\right)^{m_{ij}^h}$, where the parameters $m_{ij} = m_{ij}\left(d\right)$ and $m_{ij}^h = m_{ij}\left(d^h\right)$.

Note that the Cobb–Douglas assumption places restrictions only at the within-group level while leaving the between-group demand equations free to be arbitrarily flexible. An approximation to $M_i P_i = M_i$ can be obtained by using the observed within-group budget shares.[7] These results support a simple procedure to estimate price variation in survey data without quantity information.

2.3 Demographically varying PUVs: Practice

Given this theoretical setup, we now describe how PUVs can be obtained in practice. The description corresponds to the implementation of the `pseudounit` command.

Definition 1. PUVs—PUV($\widehat{P}_D^i$)

$$\widehat{P}_D^i = M_i P_i = M_i = \frac{1}{k_i} \prod_{j=1}^{n_i} w_{ij}^{-w_{ij}}$$

where k_i is the average of the subgroup expenditure for the ith group budget share.

5. We would like to thank an anonymous reviewer for suggesting we report how the expressions for M_i, P_i, and k_i are derived.
6. If p_{ijt} is the price in year t and p_{ij0} the price in base year 0, then $\widetilde{p}_{ijt} = p_{ijt}/p_{ij0}$ and $\widetilde{q}_{ijt} = q_{ijt}p_{ij0}$.
7. In cross-sectional data where prices are not reported or unit values cannot be derived, it is common to assume that price indexes do not vary and are therefore equal to one.

The index $\widehat{P}_D^i$ summarizes the cross-section variabilities of prices that can be added to spatially varying price indexes to resemble unit values expressed in index form as follows. In general, this technique allows the recovery of the household-specific price variability that can be found in unit values. The PUV is an index that can be compared with actual unit values after normalization, choosing the value of a specific household as a numeraire.

Definition 2. PUVs in index form—PUV($\widehat{P}_{\mathrm{DI}}^i$)

$$\widehat{P}_{\mathrm{DI}}^i = \widehat{P}_{\mathrm{D}}^i P_{rm}^i$$

where P_{rm}^i are the group-specific price indexes derived in (1).

For PUVs in index form to look like actual unit values, they have to be transformed into levels. The transformation in nominal terms is fundamental to properly capture complementary and substitution effects as shown in Atella, Menon, and Perali (2004). Cross effects would otherwise be the expression of the differential speed of change of the good-specific price indexes through time only. Note that $\widehat{P}_D^i = M_i P_i$ holds for the base year only where $\widetilde{p}_{ij} = 1$ for all regions and that time subscripts are omitted for simplicity. In subsequent time-periods,

$$M_i P_i = \frac{1}{k_i} \left(\prod_{j=1}^{n_i} w_{ij}^{-w_{ij}} \right) \left(\prod_{j=1}^{n_i} \widetilde{p}_{ij}^{w_{ij}} \right) = \widehat{P}_D^i \left(\prod_{j=1}^{n_i} \widetilde{p}_{ij}^{w_{ij}} \right)$$

which is represented by the PUV in index form, $\widehat{P}_{\mathrm{DI}}^i = \widehat{P}_D^i P_{rm}^i$. Further, $\widehat{P}_{\mathrm{DI}}^i$ will be an approximation, unless P_{rm}^i is equivalent to $\prod_{j=1}^{n_i} \widetilde{p}_{ij}^{w_{ij}}$, which resembles a Stone price index.

Definition 3. PUVs in levels—PUV($\widehat{P}_{\mathrm{DIL}}^i$)

$$\widehat{P}_{\mathrm{DIL}}^i = \widehat{P}_{\mathrm{DI}}^i \overline{y_i}$$

where $\overline{y_i}$ is the average expenditure of group i in the base year.

Early experiments with PUVs with Italian household budget data (Perali 1999, 2000; Atella, Menon, and Perali 2004; Menon and Perali 2010) and Hoderlein and Mihaleva (2008) and Berges, Pace Guerrero, and Echeverría (2012) for other datasets have provided comforting indications about the possibility of estimating regular preferences. Atella, Menon, and Perali (2004) describe the effects on the matrix of cross-price elasticities associated with several price definitions and find that the matrix of compensated elasticities is negative definite only if PUVs are used. Nominal PUVs, which more closely reproduce actual unit values, give a set of own- and cross-price effects that is more economically plausible. The derived demand systems are regular and suitable for sound welfare and tax analysis. The authors conclude that the adoption of PUVs does no harm because Lewbel's method simply consists of adding cross-sectional price variability to

aggregate price data. Therefore, Lewbel's method for constructing demographically varying prices is potentially of great practical utility.

Because goods may differ in quality from one household to another and because their associated unit values may both reflect these differences in quality, measurement errors, and endogenous expenditure information, the estimated unit values are likely to be correlated with the equation errors, and the resulting estimators will be both biased and inconsistent. The demand estimation technique should therefore account for price endogeneity by using instrumental-variable methods.

We now proceed with the description of the `pseudounit` command.

3 The pseudounit command

3.1 Syntax

The syntax of `pseudounit` is as follows:

`pseudounit` *expenditures* [*if*] [*in*] , `generate(`*varname*`)` `pindex(`*varname*`)`
[`impvars(`*varlist*`)` `seed(`#`)` `add(`#`)` `coll_rule(mean|median)` `expby(`*varname*`)`
`pdi(`*varname*`)` `year(`*varname*`)` `saving(`*filename* [, `replace`] `)`]

where *expenditures* is the list of expenditure variables of interest. The list must be specified as follows: the group expenditure first, then all the subexpenditures of the group; the `pseudounit` command verifies whether the sum of all subexpenditures sum to the group expenditure and whether each expenditure has positive or zero value.

3.2 Options

`generate(`*varname*`)` specifies the name of the new variable that will be created with the unit values associated with the group expenditure. `generate()` is required.

`pindex(`*varname*`)` specifies the variable with the price index associated with the group expenditure. For the construction of this variable, see Atella, Menon, and Perali (2004). The index must have the same base year as `expby(`*varname*`)`. `pindex()` is required.

`impvars(`*varlist*`)` specifies the variables to be used for the imputation of the zero expenditure shares. The subexpenditures must be at least two. The imputation uses the `mi impute truncreg` command, where the dependent variable is the expenditure share and the independent variables are the variables specified in `impvars(`*varlist*`)`. Without the `impvars(`*varlist*`)` option, the PUV variable is not computed for observations with subexpenditures equal to zero. If the `year()` option is specified, `year(`*varname*`)` variables are also added to `impvars(`*varlist*`)`. If the imputation with `mi impute truncreg` fails, the command switches to `mi impute pmm` using a

number of k nearest neighbors equal to 5% of the positive observations of the within-group shares. One can also use categorical variables with the appropriate syntax (see [U] **11.4.3 Factor variables**). Because imputed shares must be positive, the program checks for negative imputed values and substitutes them with the value of one. Because the procedure uses a product, this guarantees the subgroup expenditure does not contribute to the group price for that specific household.

`seed(#)` sets the random-number seed. This option is used to reproduce results. The default is `seed(159753)`.

`add(#)` specifies the number of imputations to add to the `mi` data. The total number of imputations must be comprised between 5 and 1,000. The default is `add(20)`.

`coll_rule(mean|median)` specifies the rule used to collapse `mi` data. `mi imputation truncreg` adds n replicas to the data with n imputations of the missing data, where n corresponds to the value reported in the `add(#)` option. Each dataset is identified with values of the variable `_mi_id`. These n imputations are then reduced to a single imputation by a data collapse that can be implemented with either the mean or the median. `pseudounit` executes the following command: `collapse (mean|median)` *share_var*`, by(_mi_id)`. The default statistic is the mean.

`expby(`*varname*`)` specifies the average group expenditure for the base year.[8] Without the `expby(`*varname*`)` option, the variable in the `generate(`*varname*`)` option is equal to the PUV in index form (`PUV(DI)`).

`pdi(`*varname*`)` specifies the variable generated with PUV in index form (`PUV(DI)`).

`year(`*varname*`)` specifies the name of the year variable when estimating unit values for several years. This computes the average of the subgroup expenditure for the ith group budget share for each year. The `year()` option can be used when a time series of cross-sections is available, so it is possible to compute the mean expenditure shares w_{ij} by each year.

`saving(`*filename* [`, replace`]`)` specifies the name of the disk file to be created or replaced. This option saves a kernel density graph of PUV in levels—PUV($\widehat{P}^i_{\text{DIL}}$). If the `year(`*varname*`)` option is specified, no graph is produced. If *filename* is specified without an extension, `.gph` will be assumed.

4 The pseudounit command: Examples

To become familiar with the command, the user may be interested in the following examples using `pseudounit_cmd.dta`, which is provided with the package.

8. When using one cross-section only, we can choose a given month (for instance, January) as the base year for both the price indexes and the group expenditures.

4.1 Data

For our example, expenditure data come from a series of repeated cross-sectional national household budget surveys conducted yearly by the ISTAT. Within each cross-section, households are interviewed monthly at different times during the year. The ISTAT budget survey is representative at the regional level.

The samples of household budgets for 2007 and 2008 used in this example comprise more than 23,000 households per year. To reduce the estimation burden of the present application, we have drawn a random sample of 4,935 households for 2007 and 4,916 for 2008. Household expenditures in the provided dataset have been aggregated into six groups and then transformed in budget shares: food, clothing, housing, transport and communication, education, and other goods and services.

ISTAT collects information about consumer price indexes based on the consumption habits of the whole population available monthly for each of the 106 Italian provinces with the COICOP level of disaggregation. We have chosen January 1997 as the base year. Price indexes have been matched to the two samples, accounting for the period of the year when the household was interviewed. This means that households interviewed in March have been matched with prices collected in the same month. After determining the expenditure groups, we constructed the corresponding consumer price indexes starting from the COICOP categories available for territorial disaggregation and months that have been matched to all households living in the same region and interviewed in the same month.

Table 1 reports the descriptive statistics of the price index, P^i_{rm}, of the `pseudounit` procedure for the six groups of goods and services. If users already have price information from external sources organized as in table 1, they can call the `pseudounit` procedure without following the bottom-up approach outlined above.

Table 1. Descriptive statistics of P^i_{rm} by year

year	idx_aggr1	idx_aggr2	idx_aggr3	idx_aggr4	idx_aggr5	idx_aggr6
2007	124.776	119.710	126.238	122.144	124.891	118.798
	(4.192)	(5.426)	(2.751)	(2.310)	(2.677)	(2.916)
2008	124.832	119.721	126.183	122.097	124.846	118.733
	(4.199)	(5.467)	(2.747)	(2.301)	(2.687)	(2.916)
Total	124.804	119.715	126.211	122.120	124.869	118.766
	(4.195)	(5.446)	(2.749)	(2.306)	(2.682)	(2.916)

Note: Standard errors are in parentheses.

Table 2 reports the levels of the average indexes, P^i_{rm}, by macroregion, selecting two households (HH1 and HH2) interviewed in time 1 or 2 in each macroregion to illustrate how the levels of price indexes may vary within each region by the time of the interview of the household.

Table 2. Average levels of P^i_{rm} by macroregion and households HH1 or HH2 interviewed in period 1 or 2

Macro	idx_aggr1	idx_aggr2	idx_aggr3	idx_aggr4	idx_aggr5	idx_aggr6
NW (HH1)	119.9	115.9	123.66	119.4	124.3	117.4
NW (HH2)	124.6	116.4	122.8	120.5	123.3	116.9
NE (HH1)	124.8	115.3	129.4	123.2	123.4	119.2
NE (HH2)	121.8	116.4	131.2	123.5	124.7	120.6
Centre (HH1)	125.6	123.2	121.0	121.4	130.8	116.1
Centre (HH2)	122.1	117.8	128.0	123.7	124.7	116.4
South (HH1)	123.9	109.5	123.4	113.9	114.2	117.6
South (HH2)	133.6	130.9	121.6	120.5	127.2	115.8
Islands (HH1)	122.5	116.6	125.1	123.3	124.53	121.0
Islands (HH2)	123.5	109.9	126.9	120.3	120.6	119.4

The composition of the group expenditures in our dataset is as follows:

Group expenditure 1: Food (ag6sp_1)

- Bread, cereals, and pasta (ag6sp_1_1)
- Meat, fish, and milk derivates (ag6sp_1_2)
- Fruit and vegetables (ag6sp_1_3)
- Fats and oils, sugar, alcoholic and nonalcoholic drinks and beverages, and tobacco (ag6sp_1_4)

Group expenditure 2: Clothing (ag6sp_2)

- Nonassignable clothing (ag6sp_2_1)
- Clothing and footwear: men (ag6sp_2_2)
- Clothing and footwear: women (ag6sp_2_3)
- Clothing and footwear: children (ag6sp_2_4)

Group expenditure 3: Housing (ag6sp_3)

- Rents and condominium fees (ag6sp_3_1)
- Water, energy, and heating (ag6sp_3_2)
- Home repairs and large electrical appliances (ag6sp_3_3)
- Small electrical appliances and flatware (ag6sp_3_4)

Group expenditure 4: Transport and communications (ag6sp_4)

- Private transportation (fuels and repairs) (ag6sp_4_1)
- Public transportation (ag6sp_4_2)
- Telephone (ag6sp_4_3)
- Purchase of means of transportation and telephone (ag6sp_4_4)

Group expenditure 5: Leisure and education (ag6sp_5)

- Education expenditures (ag6sp_5_1)
- Leisure (ag6sp_5_2)
- Computer, music, and television (ag6sp_5_3)
- Other (ag6sp_5_4)

Group expenditure 6: Health and other no food (ag6sp_6)

- Medical examinations and medicines (ag6sp_6_1)
- Insurance, expenditures for medical assistance, and other (ag6sp_6_2)

The dataset comprises price indexes associated with each expenditure (idx_aggr1–idx_aggr6) and mean expenditures evaluated at the base year (1997) for each of the six selected expenditure categories conditioned by region, number of household members, and month (mu_ag6sp_1–mu_ag6sp_6). Other variables are residential location, whether urban or rural (location), number of household components (nc), macroarea (ripgeo), age of the household head (etacf), education of the household head (titstucf), and the logarithm of the household total annual expenditure (lnx).

Note that in the base year, average expenditures are computed by region and month to preserve the maximum territorial and time variation.

4.2 Examples

We now implement the pseudounit command to estimate unit values for food, clothing, housing and transport, and communication to illustrate how to use the options available in the command.[9]

9. Our results are obtained using Stata 14. Possible marginal differences may be due to previous versions of Stata adopting a different pseudorandom-number generator.

The first expenditure is food (`ag6sp_1`) (subgroup expenditures are bread, cereals, and pasta [`ag6sp_1_1`]; meat, fish, milk, and other protein [`ag6sp_1_2`]; fruits and vegetables [`ag6sp_1_3`]; and fats and oils, sugar, beverages, and tobacco [`ag6sp_1_4`]).

The variables used for the imputation of zero expenditures are residential location (`location`), macroarea (`ripgeo`), number of household components (`nc`), age of the household head (`etacf`), education of the household head (`titstucf`), and the logarithm of the total annual household expenditure (`lnx`). The multiple imputation of the 0 expenditure shares generates 30 datasets that are then summarized using the mean as the default. The regional price index is `idx_aggr1_reg`, and the mean expenditure computed at the base year for food is `mu_ag6sp_1`.

The variable associated with the unit values of the food expenditure is `lwbp_aggr1`.

```
. use pseudounit_cmd.dta
. pseudounit ag6sp_1 ag6sp_1_1 ag6sp_1_2 ag6sp_1_3 ag6sp_1_4 if year==2007,
>  generate(lwbp_aggr1) pindex(idx_aggr1)
>  impvars(location i.ripgeo nc etacf i.titstucf lnx)
>  expby(mu_ag6sp_1) add(30) seed(889922) coll_rule(median)
DESCRIPTIVES STATISTICS
```

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max
PUV(D)	4971	.9405628	.9557249	.0908108	.3564937	1.09878
PUV(DI)	4971	1.173586	1.189731	.1196185	.4406378	1.478698
lwbp_aggr1	4971	516.1335	517.3575	151.9128	127.4489	1015.579

```
Note: lwbp_aggr1 is Pseudo Unit Values in Levels PUV(DIL)
```

In this case, there are no imputations because there are no zero share expenditures.

The second expenditure is clothing (subgroup expenditures are nonassignable clothing, clothing and footwear for men, clothing and footwear for women, and clothing and footwear for children).

```
. pseudounit ag6sp_2 ag6sp_2_? if year==2007,
>  generate(lwbp_aggr2) pindex(idx_aggr2)
>  impvars(location i.ripgeo nc etacf i.titstucf lnx)
>  expby(mu_ag6sp_2) add(30) seed(889922) coll_rule(median)
**** EXPENDITURE ag6sp_2_2 ****
644 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Method: truncreg regression
Limit: lower = 0
       upper = 1
Total Observations: 4971
Complete observations: 4327
Missing observations: 644
Imputed observations: 644
0 values for expenditure ag6sp_2_2 converted to 1
```

```
**** EXPENDITURE ag6sp_2_3 ****
355 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Method: truncreg regression
Limit: lower = 0
       upper = 1
Total Observations: 4971
Complete observations: 4616
Missing observations: 355
Imputed observations: 355

0 values for expenditure ag6sp_2_3 converted to 1

**** EXPENDITURE ag6sp_2_4 ****
3149 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Method: truncreg regression
Limit: lower = 0
       upper = 1
Total Observations: 4971
Complete observations: 1822
Missing observations: 3149
Imputed observations: 3149

0 values for expenditure ag6sp_2_4 converted to 1

DESCRIPTIVES STATISTICS
```

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max
PUV(D)	4971	1.049126	1.051407	.0675399	.4734136	1.181265
PUV(DI)	4971	1.25613	1.259243	.1018521	.6206452	1.556212
lwbp_aggr2	4971	174.4161	177.3384	67.20049	24.84464	555.8301

```
Note: lwbp_aggr2 is Pseudo Unit Values in Levels PUV(DIL)
```

In this case, there are no imputations for subgroup expenditure `ag6sp_2_1`, but there are imputations for subgroup expenditures `ag6sp_2_2`, `ag6sp_2_3`, and `ag6sp_2_4`.

The third expenditure is housing (subgroup expenditures are rent and condo expenses; water, energy, and heating; home repairs and large electrical appliances; small electrical appliances and flatware). The `lwbp_aggr3` variable is created for each year (`year`).

```
. pseudounit ag6sp_3 ag6sp_3_?,
>  generate(lwbp_aggr3) pindex(idx_aggr3)
>  impvars(location i.ripgeo nc etacf i.titstucf lnx)
>  expby(mu_ag6sp_3) year(year)

**** EXPENDITURE ag6sp_3_3 ****
1668 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Method: truncreg regression
Limit: lower = 0
       upper = 1
Total Observations: 9859
Complete observations: 8191
Missing observations: 1668
Imputed observations: 1668

0 values for expenditure ag6sp_3_3 converted to 1
```

DESCRIPTIVES STATISTICS

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max
2007						
PUV(D)	4971	.9480497	.931581	.1956049	.4917437	1.509897
PUV(DI)	4971	1.196825	1.177358	.2483218	.6092809	1.940222
lwbp_aggr3	4971	729.4016	698.4557	248.5652	160.168	1894.283
2008						
PUV(D)	4888	.9482172	.9302261	.1977599	.4697024	1.527056
PUV(DI)	4888	1.196471	1.173393	.2506707	.5757318	1.938348
lwbp_aggr3	4888	723.3252	692.9372	246.5752	187.4321	1780.587

Note: lwbp_aggr3 is Pseudo Unit Values in Levels PUV(DIL)

The fourth expenditure is transport and communications (subgroup expenditures are private and public transportation, telephone, and purchase of transportation means).

```
. pseudounit ag6sp_4 ag6sp_4_? if year==2008,
>  generate(lwbp_aggr4) pindex(idx_aggr4)
>  impvars(location i.ripgeo nc etacf i.titstucf lnx)
>  expby(mu_ag6sp_4) seed(889922) coll_rule(median) saving(kd_sp4, replace)
**** EXPENDITURE ag6sp_4_1 ****
907 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Method: truncreg regression
Limit: lower = 0
       upper = 1
Total Observations: 4888
Complete observations: 3981
Missing observations: 907
Imputed observations: 907
0 values for expenditure ag6sp_4_1 converted to 1
**** EXPENDITURE ag6sp_4_2 ****
379 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Method: truncreg regression
Limit: lower = 0
       upper = 1
Total Observations: 4888
Complete observations: 4509
Missing observations: 379
Imputed observations: 379
0 values for expenditure ag6sp_4_2 converted to 1
**** EXPENDITURE ag6sp_4_3 ****
60 observations to impute
MULTIPLE IMPUTATION OVERVIEW
Pay attention: imputation method switched to pmm
Method: pmm regression
Total Observations: 4888
Complete observations: 4828
Missing observations: 60
Imputed observations: 60
Number of k nearest neighbors: 241
0 values for expenditure ag6sp_4_3 converted to 1
```

```
DESCRIPTIVES STATISTICS
```

Variable	Obs.	Mean	Median	Std. Dev.	Min	Max
PUV(D)	4888	.8136841	.8143076	.1388408	.4111537	1.281451
PUV(DI)	4888	.9933574	.9960147	.169638	.5019528	1.561017
lwbp_aggr4	4888	360.8287	319.7612	223.4753	10.74425	2049.023

```
Note: lwbp_aggr4 is Pseudo Unit Values in Levels PUV(DIL)
(file kd_sp4.gph saved)
```

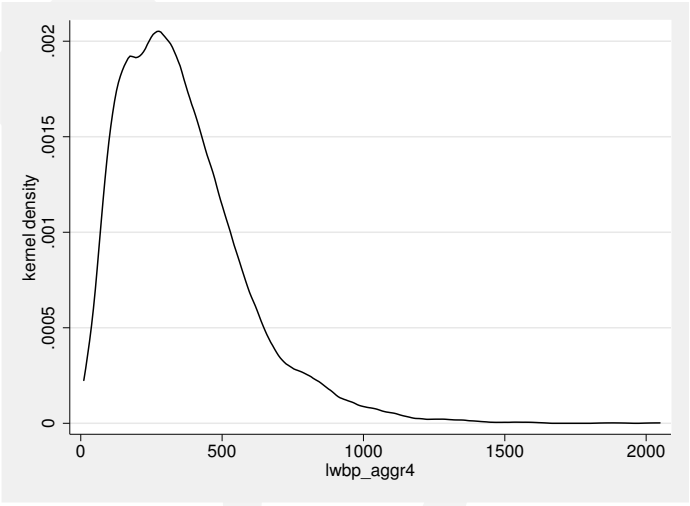

Figure 1. `lwbp_aggr4` kernel density estimation

Note that the multiple-imputation procedure using the `truncreg` method for the subgroup expenditure `ag6sp_4_3` failed and that the program switched to the predictive mean matching method.

5 Conclusions

The main objective of the `pseudounit` command presented here is to make household budget surveys that collect only information about expenditures suitable for demand and welfare analysis. Thanks to the `pseudounit` command, the lack of information about quantities no longer precludes the possibility of deriving household-specific prices (unit values) and of estimating complete demand systems suitable for welfare analysis.

6 Acknowledgments

The authors wish to thank an anonymous referee and Lucia Echeverrìa for helpful comments and suggestions. Any errors and omissions are the sole responsibility of the authors.

7 References

Atella, V., M. Menon, and F. Perali. 2004. Estimation of unit values in cross sections without quantity information and implications for demand and welfare analysis. In *Household Behaviour, Equivalence Scales, Welfare and Poverty*, ed. C. Dagum and G. Ferrari, 195–220. Heidelberg: Physica-Verlag.

Barten, A. P. 1964. Family composition, prices and expenditure patterns. In *Econometric Analysis for National Economic Planning*, ed. P. Hart, G. Mills, and J. K. Whitaker. London: Butterworth.

Berges, M., I. Pace Guerrero, and L. Echeverría. 2012. La Utilización de Precios Implícitos o de Pseudo Precios Implícitos en la Estimación de un Sistema de Demandas QUAIDS para Alimentos. Nülan. Deposited Documents 1675, Centro de Documentación, Facultad de Ciencias Económicas y Sociales, Universidad Nacional de Mar del Plata.

Coondoo, D., A. Majumder, and R. Ray. 2004. A method of calculating regional consumer price differentials with illustrative evidence from India. *Review of Income and Wealth* 50: 51–68.

Dagsvik, J. K., and L. Brubakk. 1998. Price indexes for elementary aggregates derived from behavioral assumptions. Discussion Paper No. 234, Statistics Norway, Research department.

Deaton, A. 1987. Estimation of own- and cross-price elasticities from household survey data. *Journal of Econometrics* 36: 7–30.

———. 1989. Household survey data and pricing policies in developing countries. *World Bank Economic Review* 3: 183–210.

———. 1990. Price elasticities from survey data: Extensions and Indonesian results. *Journal of Econometrics* 44: 281–309.

———. 1998. Getting prices right: What should be done? *Journal of Economic Perspectives* 12: 37–46.

Deaton, A. S. 1988. Quality, quantity, and spatial variation of price. *American Economic Review* 78: 418–430.

Deaton, A. S., and J. Muellbauer. 1980. *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

Frisch, R. 1959. A complete scheme for computing all direct and cross demand elasticities in a model with many sectors. *Econometrica* 27: 177–196.

Hoderlein, S., and S. Mihaleva. 2008. Increasing the price variation in a repeated cross section. *Journal of Econometrics* 147: 316–325.

Lahatte, A., R. Miquel, F. Laisney, and I. Preston. 1998. Demand systems with unit values: A comparison of two specifications. *Economics Letters* 58: 281–290.

Lecocq, S., and J.-M. Robin. 2015. Estimating almost-ideal demand systems with endogenous regressors. *Stata Journal* 15: 554–573.

Lewbel, A. 1989. Identification and estimation of equivalence scales under weak separability. *Review of Economic Studies* 56: 311–316.

Menon, M., and F. Perali. 2010. Econometric identification of the cost of maintaining a child. In *Research on Economic Inequality, Volume 18: Studies in Applied Welfare Analysis: Papers from the Third ECINEQ Meeting*, ed. J. A. Bishop, 219–256. Bingley, UK: Emerald.

Perali, F. 1999. Stima delle scale di equivalenza utilizzando i bilanci familiari ISTAT 1985–1994. *Rivista Internazionale di Scienze Sociali* 107: 481–541.

———. 2000. Analisi di Tassazione Ottimale Applicata al Consumo di Bevande. In *Microeconomia Applicata, vol. I*, ed. F. Perali. Roma: Carocci.

———. 2003. *The Behavioral and Welfare Analysis of Consumption: The Cost of Children, Equity and Poverty in Colombia.* Dordrecht: Kluwer.

Slesnick, D. T. 1998. Empirical approaches to the measurement of welfare. *Journal of Economic Literature* 36: 2108–2165.

About the authors

Martina Menon is an assistant professor of economics at the University of Verona in Italy. Her main research interests are economics of the family, consumption analysis, welfare analysis, and applied econometrics.

Federico Perali is a full professor of economics at the University of Verona in Italy. His main research interests are economics of the family, consumption analysis, welfare analysis, and applied econometrics.

Nicola Tommasi is a research assistant at the University of Verona in Italy. His main research interests are applied econometrics and statistical programming.