Proceedings of Q2006
European Conference on Quality in Survey Statistics

Estimating real income at regional level using a CPI bias correction

Rosa Bernardini Papalia[1]

1. Introduction

The regional Consumer Price Index (CPI) is an issue of great interest to measure the price movements for a specific local area of interest and can be used: (i) to deflate monetary measures of household living standards, and (ii) to update poverty lines, at a regional level. The issue that consumer price index bias may affect the measurement of the trend rate of poverty reduction is relevant also to countries where either the consumer price index or a variant of it, such as a price index of low-income workers, is used to update poverty lines (Deaton and Tarozzi 2000; Deaton, 2003).
In literature, much attention has been devoted to the nature and extent of different sources of bias in consumer price index compilation based on a modified Laspeyres index (Diewert, 1998; Balk 1999), and to the measurement of each particular type of bias. This type of index is known to produce a number of biases, compared to the conceptual standard of a true cost of living index (Hausman, 2003): substitution (or formula) bias; elementary index bias; outlet substitution bias; new goods and quality adjustment or linking bias. In particular, because consumers may substitute away from higher priced goods and outlets, while a Laspeyres index continues measuring the price of the higher priced items (from the original selected outlets), the consumer price index will be an upwardly biased estimate of changes in the true cost of living. In this study we introduce a method for measuring consumer price index bias at regional level that uses Engel's law (Hamilton, 2001).
The idea is to calculate consumer price index bias by estimating the income elasticity of food using cross-sectional micro data and to employ these estimates to measure the increase in households' real income over time by controlling for changes in relative prices and in demographic characteristics at regional level. More specifically, given that food's budget share is inversely related to household real income, by controlling for movements in relative prices and households characteristics, it is possible to infer changes in real incomes from movements in the share of food. This approach gives reduced form estimates of the overall bias in the consumer price index, inferred from movements in food Engel curves over time.

2. Estimating framework for CPI bias measurement at regional level

The Laspeyres consumer price index which finds the cost of purchasing a fixed basket in a base period and the cost of buying the same basket in the present, is

[1] Department of Statistics, University of Bologna, Italy (e-mail:bernardini@stat.unibo.it).

known to produce a number of biases, such as *outlet bias*, *quality change* and *new products bias*.
Following the work of Hamilton (2001) we show how to use different time periods of cross-sectional micro data to identify CPI bias at regional level and to infer CPI bias from a food Engel curve.
The advantage of food as an indicator good is that its low income elasticity makes its budget share sensitive to the measurement of income, whereas goods with income elasticities close to one will have budget shares that are unchanged through time even if income growth is mismeasured. Food is also a non-durable, implying that expenditures in one period cannot provide a flow of consumption in another, and is likely to be separable from other goods in consumers' utility functions.

We start by introducing the Leser-Working form of the Engel model:

$$w_{i,j,t} = \phi + \gamma(\ln P_{F,j,t} - \ln P_{NF,j,t}) + \beta(\ln Y_{i,j,t} - \ln P_{j,t}) + X'\theta + u_{i,j.t} \quad (1)$$

where $w_{i,j,t}$ is the budget share of food for household *i* in region *j* and time period t, $P_{F,j,t}$, $P_{NF,j,t}$, and $P_{j,t}$ represent the true but unobserved prices of food, non-food, and all goods, Y is the household's total income (which is measured by total expenditure), X is a vector of individual household characteristics and u is the residual.
The true cost of living is treated as a geometric weighted average of the prices of food and non-food:

$$\ln P_{j,t} = \alpha \ln P_{F,j,t} + (1-\alpha)\ln P_{NF,j,t} \quad (2)$$

and it is also assumed that all prices of a good G (either food, non-food, or all goods) are measured with error:

$$\ln P_{G,j,t} = \ln P_{G,j,0} + \ln(1+\Pi_{G,j,t}) + \ln(1+E_{G,t}) \quad (3)$$

where $\prod_{G,j,t}$ represents the cumulative percentage increase in the CPI-measured price of good G from period 0 to period t and $E_{G,t}$ is the period-t percent cumulative measurement error in the cost-of-living index since the base period.
By substituting equation (3) into (2), we obtain:

$$\ln(1+E_t) = \alpha \ln(1+E_{F,t}) + (1-\alpha)\ln(1+E_{NF,t}) \quad (4)$$

Assuming that CPI bias does not vary geographically, substituting equations (2), (3) and (4) into equation (1) gives:

$$\begin{aligned} w_{i,j,t} = {} & \phi + \gamma\left[\ln(1+\Pi_{F,j,t}) - \ln(1+\Pi_{NF,j,t})\right] + \\ & + \beta(\left[\ln Y_{i,j,t} - \ln(1+\Pi_{j,t})\right] + X'\theta + \\ & + \gamma\left[\ln(1+E_{F,t}) - \ln(1+E_{NF,t})\right] - \beta\ln(1+E_t) + \\ & + \gamma\left[\ln P_{F,j,0} - \ln P_{NF,j,0}\right] - \beta\ln P_{j,0} + u_{i,j,t}. \end{aligned} \quad (5)$$

By using a cross-section/time series database with micro data on income and food expenditure as well as other variables such as family composition which would influence food expenditure, as well as cross-section CPI for all consumption, for food and non-food, over the entire data period for a single region, the following empirical version of equation (5) can be estimated:

$$\begin{aligned} w_{i,j,t} &= \hat{\phi} + \gamma\left[\ln(1+\Pi_{F,j,t}) - \ln(1+\Pi_{NF,j,t})\right] + \\ &+ \beta(\left[\ln Y_{i,j,t} - \ln(1+\Pi_{j,t})\right] + X'\theta + \\ &+ \sum_{t=1}^{T} \delta_t D_t + u_{i,j,t.} \end{aligned} \tag{6}$$

where D_t is a dummy variable equal to 1 in period t, δ_t is its coefficient, and $\hat{\phi}$ is the intercept from equation (5), plus the coefficients of the omitted time dummies.
The key role of the time dummy variables to the measurement of CPI bias emerges because:

$$\delta_t = \gamma\left[\ln(1+E_{F,t}) - \ln(1+E_{NF,t})\right] - \beta \ln(1+E_t)\ . \tag{7}$$

Under the assumption that the relative bias between food and non-food is constant across time periods and writing the previous expression in terms of the cumulative bias in the CPI for all goods, then we have:

$$\ln(1+E_t) = \frac{\delta_t}{-\beta - \dfrac{\gamma(1-r)}{1-\alpha(1-r)}}\ . \tag{8}$$

It follows that the bias can be identified up to an unknown parameter, r, which is the ratio of the CPI bias in food and non-food, and also depends on α, which is foods' share in the cost-of-living index. The equation (8) can be reduced to:

$$\ln(1+E_t) \approx -\frac{\delta_t}{\beta}\ . \tag{9}$$

if either γ or (1-r) is close to zero. If r<1 as seems plausible – food is less badly biased than nonfood - then (9) understates the bias, increasingly as r falls increasingly below unity. Thus, a lower bound for cumulative percentage CPI bias at period t is given by a simple ratio of estimated coefficients from equation (6):

$$1 - \exp\left\{-\frac{\delta_t}{\beta}\right\}. \tag{10}$$

If the assumptions underlying equation (9) are satisfied, the cumulative CPI bias is found by dividing the coefficient on the dummy variable for the round by the income coefficient; the average monthly bias can be found by dividing the difference between cumulative bias estimates by the number of months separating them.

To identify the parameter on food prices, γ, it is possible to use cross-sectional variations in inflation rates, rather than price levels. Nevertheless, this period by period variation in an aggregate price index for food relative non-food is perfectly correlated to the time dummy variables, D_t, and an appropriate estimator has to be used. To this end, a Generalized Maximum Entropy (GME) estimation approach (Bernardini Papalia 2004; Golan et al. 1996) is here suggested. The maximum entropy-based estimators are most efficient relative to traditional estimators in particular when data constraints for each observation are included in the maximum entropy-based problem formulations. Second, they are able to produce estimates in models where the number of parameters exceeds the number of data points and in models characterized by a non-scalar identity covariance matrix. Third, prior information can be introduced by adding suitable constraints in the formulation without imposing strong distributional assumptions.

Without geographic variation in the price of food, equation (6) becomes:

$$w_{i,t} = \hat{\phi} + \beta(\left[\ln Y_{i,t} - \ln(1+\Pi_t)\right] + X'\theta + \\ + \sum_{t=1}^{T} \delta_t D_t + u_{i,j,t}. \tag{11}$$

This model specification can be used when cross-sectional variation in food prices is unavailable. Here, the dummy variables measure not just the CPI bias of equation (7) but also the effect on budget shares of intertemporal variation in the measured inflation rate for food relative to non-food. Hence, the cumulative percentage CPI bias at time t is calculated from:

$$1 - \exp\left\{ - \frac{\delta_t - \bar{\gamma}\left[\ln(1+\pi_{F,t}) - \ln(1+\pi_{NF,t})\right]}{\beta} \right\} \tag{12}$$

where $\bar{\gamma}$ has to be computed from outside of the estimated parameters for equation (11).

Using regionally disaggregated data for the food and nonfood inflation rates, equations (6) and (9) provide the basic framework, following the approach of Hamilton (2001) of employing food and nonfood inflation rates rather than price levels to identify $\bar{\gamma}$.

3. An empirical application

3.1 Estimation of regional price indexes

The Laspeyres CPI is defined as a fixed-quantity price index that measures the price change in a fixed market basket of consumption goods and services that are purchased by the reference population (Turvey, 2002).

Let $\hat{e}_g^0$ be an estimator of the expenditure on commodity group *g* ($\hat{e}^0 = \sum_g \hat{e}_g^0$); the Laspeyres CPI at time *t* with base year *0* is given by:

$$I^{t}=\sum_{g} I_{g}^{t} w_{g}^{0}=\sum_{g} I_{g}^{t} \hat{e}_{g}^{0} / \hat{e}^{0}$$

$$\sum_{g} w_{g}^{0}=1,$$

$$w_{g}^{0}=\frac{\sum_{i \in H^{0}} e_{i}^{0} w_{ig}^{0}}{\sum_{i \in H^{0}} e_{i}^{0}}, \qquad (13)$$

$$e_{i}^{0}=\sum_{g} e_{ig}^{0}$$

where: I^t_g is the price index of commodity group (or expenditure item) *g*, (*g=1,..,G*), e^0_i denotes household *i* total expenditures (i=1,...,H), w^0_g is the budget shares for expenditure category g in the CPI, and w^0_{ig} denotes household *i* budget share for expenditure category *g*. The budget shares for expenditure category *g* in the CPI, w^0_g, equals the share of the expenditure on *g* in the total base year expenditure on all commodities for a specific group of households.

Expenditure shares for each good are treated as if they were those of an aggregate “super-household” representative of the specific population group. More specifically, the CPIs use weights which reflect the composition of the estimate aggregate values of the reference population. Each household contributes to these weights by an amount proportional to its expenditure. Such weighting is named “plutocratic” in contrast with the “democratic” type of weighting which gives equal importance to all households by averaging consumption value proportions over the whole reference population. The aggregate CPI is computed with the weights which reflect the expenditure of an average household. The resulting Laspeyres price index, known as the plutocratic Laspeyres price index (Prais, 1959), measures the price change of the total base period consumption and it may be interpreted as the price change of the “representative” household’s base period consumption relative to the reference population group.

The regional index is computed by calculating: *(i)* city indices of product; *(ii)* regional indices of product, as weighted average of the elementary city indices of product with weights equal to the population in the cities; *(iii)* regional index, as weighted average of the regional indices of product with weights equal to the expenditures in the region. Weights are obtained from a Family Budget Survey and from other sources in combination with National Accounts data on regional household consumption within the Italian economic territory.

In this view, the regional consumer price index is calculated for the Umbria region by using monthly data over the period January to June 2001 with regard to the 2000 basket and prices, and by assuming the month of December 2000 as the time base. The price indices for each product in four different cities (Perugia, Terni, Città di Castello, Orvieto) are combined in the regional price index for product using weights which represent the relative importance of the expenditure in each city and since expenditure data are not available at the city level, population weights are used (Bernardini Papalia 2004). Then, the price indices for all products are combined to obtain the regional index using weights which represent the relative importance of the expenditure for each product in the region. The data base used to estimate the expenditure shares for subgroups of the reference population is derived from the Italian Expenditures Survey and contains detailed information about the expenditure, together with a great number of household characteristics of a sample of households in 2001.

The results of the computation of the Consumer Price Index for the Umbria region are presented in Table 1. From January to June 2001 the CPI for the Umbria region rose by 1.9%; it rose at a slower pace than the Italian CPI and it was above the Italian CPI for the entire period; the Italian CPI seems to have escalated at a faster rate than the RCPI for all the main consumption groups. Over the entire period, the greatest price increases at the main component level were in Food and Beverages, Alcohol and Tobacco, Transportation, Restaurant and Hotels, while the greatest declines were for Housing and Communication.

Tab. 1 - Consumer Price Index of the Umbria region (December 2000=100, January - June 2001)

Expenditure Category	*Jan-01*	*Feb-01*	*Mar-01*	*Apr-01*	*May-01*	*Jun-01*
Food and Beverages	101.1	101.6	102.1	102.6	103.3	103.7
Alcohol and Tobacco	100.1	100.1	100.3	103.2	103.2	103.3
Clothing and Footwear	100.0	100.0	100.3	100.7	100.9	100.9
Housing	100.1	99.7	99.9	100.1	98.8	98.7
Household furnishings	100.1	100.6	100.7	100.8	101.8	101.8
Medical care	100.6	100.6	100.6	100.7	100.7	100.7
Transportation	99.2	99.4	99.5	99.9	101.0	101.2
Communication	99.5	99.2	99.1	99.0	98.7	98.7
Recreation	102.0	102.0	102.0	102.1	102.3	102.2
Education	100.0	100.0	100.0	100.0	100.0	100.0
Restaurants and Hotels	102.7	103.2	103.5	105.1	105.2	105.4
Other goods and services	101.1	101.2	101.4	102.0	102.1	102.1
All items	100.6	100.8	101.0	101.5	101.8	101.9

The consumption expenditure of some special groups of the reference population: *(i)* low-income household (FamL), *(ii)* high-income household (FamH), *(iii)* one-person type family (Single), and *(iv)* family without children (No-kids), is quite different from the average expenditure of the general population (see Tab. 2).

Tab. 2 - Distribution of total average expenditures for the major components of the Consumer Price Index (Umbria region. Relative share in percentage. Household Survey 2000)

Expenditure Category	*W Umbria*	*W H-F*	*W L-F*	*W Nokids*	*W Single*
Food and Beverages	18.0	16.1	23.1	18.4	17.4
Alcohol and Tobacco	1.6	1.5	1.9	1.6	1.5
Clothing and Footwear	7.6	8.3	5.6	7.5	6.7
Housing	27.8	24.2	38.2	29.5	34.9
Household furnishings	7.8	9.5	3.0	7.0	4.5
Medical care	3.5	3.4	3.7	3.9	4.8
Transportation	14.6	15.8	11.1	13.8	11.2
Communication	2.6	2.4	3.1	2.6	3.1
Recreation	5.3	6.0	3.4	5.2	5.5
Education	0.3	0.3	0.2	0.2	0.3
Restaurants and Hotels	2.4	2.7	1.7	2.3	3.4
Other goods and services	8.6	9.7	5.2	8.0	6.8
All items	100	100	100	100	100

More specifically, an examination of Table 2 shows that, at the level of Housing, Communication, and Medical care, the total average expenditure relative to some sub-groups of the reference population is higher than that of the reference population for low-income households, no-kids and single type families. In the case of high-income households, the relative share of total expenditures for the Clothing, Furnishings and Transportation categories are 8.3, 9.5 and 15.8 percent as compared to 7.6, 7.8 and 14.6 percent for the reference population and were 9.3%, 22% and 8.2% higher. Low-income households' average expenditures on Food, Housing, and Medical care categories were 23.1, 38.2 and 3.7 percent as compared to 18, 27.8 and 3.49 percent for the reference population and were 28.5%, 37.2% and 6.0% higher.

3.2. Estimation of CPI bias at regional level

The CPI bias is estimated from the food regression using a sample of household of the Umbria region according to the equation (6). A Generalized maximum entropy estimator is used to deal with the collinearity problems (Bernardini Papalia, 2004).

The dependent variable used is the share of expenditures devoted to all food; control variables include the real total expenditures, relative price changes, demographic characteristics, time dummies. Using a more limited set of demographic controls does not affect the bias estimates, so we include only the household size neglecting control variables relative to the work status of the head household, number of children, etc. Augmenting the model with a quadratic expenditure term does not change the dummy variable coefficients showing the downward drift in the food Engel curve, and in fact the quadratic terms are not statistically significant in the estimation results.

The regression results yield reasonable estimates of expenditures and price elasticities; control variables have the expected sign. Food and non-food inflation rates at regional level are computed according to the methodology we presented in section 3.1. Table 3 summarizes cumulative bias estimates and presents estimates corrected for relative price changes.

Tab. 3 – Estimates of cumulative CPI bias in Umbria region, (January - June 2001)

	Bias – estimates adjusted for relative price changes
Jan-01	0.809
Feb-01	0.923
Mar-01	0.000
Apr-01	0.891
May-01	0.838
June-01	0.000

Some differences emerged when the mean family income is deflated using the arithmetic mean over six months of the regional CPI index with and without the bias adjustment. For example, deflating income with the regional CPI yields a real income of 44381. When the bias-adjusted regional CPI index is used to deflate mean family expenditure yields a real income of 44636 euros. The result is a -254 difference in purchasing power, or approximately a -0.6 percent difference. This evidence emerged in analyzing all specific sub-groups of the reference population.

4. Conclusions

In this paper, following the work of Hamilton (2001) the idea is to compute a consumer price index bias by estimating the income elasticity of food using cross-sectional micro data at regional level. This approach gives reduced form estimates of the overall bias in the regional consumer price index, inferred from movements in food Engel curves over time.

We have estimated Engel functions for the food budget share of a sample of households living in a specific region of Italy, based on data from January to June 2001 from the Italian Family Budget Survey.

We find an average CPI bias relative to the Umbria region of about two percentage points per month during the period we analyze.

The cumulative effect of this bias causes an understatement of the growth performance of this region. We find that, the level of real percapita GDP in 2001 may be undersated by up to 0.6% compared with using a bias-corrected deflator.

References

Bernardini Papalia R., (2004) Estimating item weights of consumer price indexes for small reference populations, in: Proceedings of the "European Conference on Quality and Methodology in Official Statistics (Q2004)", Mainz.

Deaton, A. (2003), "Prices and Poverty in India, 1987-2000", *Economic and Political Weekly*, 25, pp. 362-368.

Deaton, A. and Tarozzi, A. (2000), *Prices and Poverty in India*, Mimeo, Princeton University.

Diewert, W.E. (1998), "Index Number Issues in the Consumer price Index", *Journal of Economic Perspectives,* 12,1, pp. 47-58.

Golan A., Judge G., and D. Miller (1996), Maximum entropy econometrics: robust estimation with limited data, Wiley.

Hausman, J. (2003), "Sources of Bias and Solutions to Bias in the Consumer Price Index", *Journal of Economic Perspectives,* 17,1, pp. 23-44.

Hamilton, B. (2001), "Using Engel's Law to Estimate CPI bias", *American Economic Review*, 91, 3, 619-630.

Prais S. (1959), "Whose Cost of Living?", *Review of Economic Studies,* 26, 126-134.

Turvey R. (2002), CPI Manual. True Cost of Living Indexes.