# What do we Learn from Recall Consumption Data?

*Erich Battistin*
*Raffaele Miniaci*
*Guglielmo Weber*

THE INSTITUTE FOR FISCAL STUDIES
WP 00/10

# What Do We Learn from Recall Consumption Data?

Erich Battistin* Raffaele Miniaci† Guglielmo Weber‡

10 May 2000

**Abstract**

In this paper we use two complementary Italian data sources (the 1995 Istat and Bank of Italy household surveys) to generate household-specific non-durable expenditure and savings measures in the Bank of Italy sample that contains relatively high-quality income data. We show that food expenditure data are of comparable quality and informational content across the two surveys, once heaping, rounding and time averaging are properly accounted for. We therefore depart from standard practice and rely on structural estimation of an inverse Engel curve on Istat data to impute non-durable and total expenditure to Bank of Italy observations, and show how these estimates can be used to analyse saving and consumption age profiles conditional on demographics.

**Acknowledgments** We are grateful for helpful discussions with Enrico Rettore, Nicoletta Rosati and particularly Jean-Marc Robin, and for comments by audiences at ESEM99, UCL, UCY, Università di Padova and INSEE. We would like to thank Viviana Egidi and Giuliana Coccia from Istat, and Giovanni D'Alessio from the Bank of Italy for making the data available to us. This research was financed by CNR and MURST and sponsored by the ISTAT workgroup exploring the feasibility of constructing an integrated data bank on household consumption and income from ISTAT and Bank of Italy survey information.

*Department of Statistics, Padua University, and IFS
†Department of Economics, Padua University
‡Department of Economics, Padua University, IFS and CEPR

## Executive Summary

Survey data on household behaviour should ideally cover a number of different areas: labour supply by individual household members, income, wealth, and expenditure on various items. Expenditure data is particularly hard to collect: the best information is based on diaries filled by participants over a period of time (as in the British Family Expenditure Survey). Filling diaries is a time consuming task, though. For this reason in many surveys recall questions are asked instead (as in the U.S. Panel Study of Income Dynamics).

In this paper we address the issue of the informational content of recall consumption questions. We use two Italian surveys (the Survey on Household Income and Wealth run by Bank of Italy and the Survey on Family Budgets run by the Italian Statistical Office): in the former recall questions are asked on total spending on non-durable goods and services and on food expenditure; in the latter detailed diary records are collected. We then compare histograms for both types of consumption measures: in both cases, recall questions are obviously affected by major heaping and rounding problems.

When we explicitly model the recall error process, we find that the underlying distribution of true expenditure is the same across surveys for food, but is different for total non-durable expenditure. We interpret this as evidence that the recall food consumption question is informative and can be used in estimation, once heaping and rounding errors are taken into account. This involves imputing a new measure of food consumption for each observation.

We then address another issue: Can the two surveys be used in conjunction to generate predictions for non-durable consumption in the recall-based survey? The standard answer to this question requires using information on a common set of explanatory variables (such as family composition, education, region of residence, age etc.). A regression of the variable of interest (non-durable consumption) on these variables should be run on the diary-based survey and used to predict in the other survey. In our case, there is information to be used on food consumption in the recall-based survey. Imputed food consumption should in principle be a useful variable to predict total non-durable consumption. We show how the prediction exercise can be extended to cover this case, taking into account that food is jointly determined with non-durable consumption.

We finally compare predictions from both methods and show that even though their statistical properties are similar, their economic implications are markedly different.

## 1. Introduction

In this paper we use two complementary Italian data sources (the 1995 Istat Survey on Family Budgets -SFB - and the 1995 Bank of Italy Survey on Household Income and Wealth- SHIW) to generate household-specific expenditure (non-durable and total) and savings measures in the Bank of Italy SHIW sample that contains relatively high-quality income data, but relatively low-quality, recall-based expenditure data.

We show that food expenditure data are of comparable quality and informational content across the two surveys, once heaping, rounding and time averaging are properly accounted for. For other expenditure definitions (non-durable and total) there are major differences across the two surveys. We make the identifying assumption that the diary-based expenditure data provided in the Istat survey is at worst affected by classical (zero-mean) measurement error.

As emphasized in the econometric literature (Arellano and Meghir, 1992), matching data sets is problem-specific. In our case, the problem is to use detailed expenditure information from a household survey to impute it into another survey that has better income and wealth data. Our aim is to construct economically meaningful definitions of savings at the household level - these will be used to generate mean and median age profiles controlling for key demographic variables. Given the availability of reliable expenditure information in one of the two surveys, however, we can depart from the standard complementary data sets practice of reduced form estimation and rely on structural estimation of an inverse Engel curve on Istat SFB data to impute non-durable and total expenditure to Bank of Italy SHIW observations.

The paper is organized as follows. Section 2 provides a description of the key features of the two surveys and highlights the need for a reliable non-durable expenditure measure in SHIW. Section 3 introduces our Engel-curve-based data matching approach and discusses some statistical methods that we use: Rosenbaum and Rubin's propensity score adjustment as well as Heijtan and Rubin's coarse data correction procedure. Section 4 provides evidence on the presence of heaping and rounding problems in SHIW and on the comparability of food and non-durable expenditure information across the two surveys, once heaping and rounding have been taken into account. In Section 5 we use non-parametric and semi-parametric techniques to establish that a linear double-log specification is an adequate representation for the inverse Engel curve in the SFB data. We also present parametric estimates that confirm the importance of allowing for simultaneity in estimation. Section 7 describes the prediction/matching exercise and shows how its results change the saving rate age profile in SHIW.

Table 2.1: Descriptive Statistics for SHIW (N=8145)

| **Variables** | **Median** | **Mean** | **Std. Dev.** | **Min** | **Max** |
|---|---|---|---|---|---|
| Food | 800 | 852.0364 | 443.7052 | 0.0 | 5000 |
| Non-durable | 1600 | 1828.841 | 946.6816 | 100 | 10000 |
| Total | 2400 | 2782.805 | 1707.076 | 135 | 35966.67 |
| Income | 3100 | 3647.452 | 2897.274 | -5666.667 | 64256.42 |
| Head age | 54 | 54.2334 | 15.1689 | 17 | 94 |
| # Members | 3 | 2.9408 | 1.3508 | 1 | 9 |
| Prop. children | 0.3333 | 0.2752 | 0.2411 | 0 | 0.8571 |
| Prop. over 60 | 0 | 0.3230 | 0.4141 | 0 | 1 |

## 2. Data Description

The two major sources of information on household income and consumption in Italy are the heavily used Bank of Italy Survey of Household Income and Wealth (SHIW- documented in Brandolini and Cannari, 1994) and the recently released ISTAT Survey on Family Budgets (SFB). The former has been run every second year since 1987, whereas the latter is run every year (and is available to researchers from 1985). A comparison of the income data is in Brandolini (1998), that suggests that better quality data are to found in SHIW.

As far as consumption is concerned, the Istat SFB follows the standard international procedure of exploiting both information from recall questions for more durable items bought in the quarter prior to the interview and diary-based records of purchases carried out within a twenty-day period. The SHIW instead contains questions on purchases of specific durable items, and asks the average monthly expenditure on food and on non-durable items (excluding rent and housing maintenance) over the previous year. In this paper we want to improve on (recall question-based) consumption information in SHIW by using diary-based information from SFB.

We gained access to the fully disaggregate version of the SFB and were able to construct expenditure items in SFB that are fully comparable to the Bank of Italy definition used in SHIW. We also re-coded a number of relevant variables to make definitions comparable across the two surveys (see Rosati (1998) for further details).

In Tables 2.1, 2.2 and 2.3 we present descriptive statistics for the two samples (SHIW for Bank of Italy; SFB for Istat). Similar tables can be obtained if sampling weights are used, but results are very close.

A first comparison based on tables 2.1, 2.2 reveals minor discrepancies for food

Table 2.2: Descriptive Statistics for SFB (N=33143)

| **Variables** | **Median** | **Mean** | **Std. Dev.** | **Min** | **Max** |
|---|---|---|---|---|---|
| Food | 734.553 | 825.0641 | 489.7272 | 8.4 | 7075.75 |
| Non-durable | 2068.558 | 2561.713 | 1996.311 | 82.248 | 24760.99 |
| Total | 2616.9 | 3165.427 | 2265.834 | 179.855 | 30588.18 |
| Income | 2975 | 3506.085 | 1966.703 | 300 | 19919.25 |
| Head age | 53 | 53.7652 | 15.9067 | 14 | 99 |
| # Members | 3 | 2.7988 | 1.3193 | 1 | 10 |
| Prop. children | 0.3333 | 0.2556 | 0.2381 | 0 | 0.875 |
| Prop. over 60 | 0 | 0.3083 | 0.4158 | 0 | 1 |

Table 2.3: Comparisons of key indicators

| **Area** | **SFB** | **SHIW** |
|---|---|---|
| Northern Italy | 0.4423 | 0.4433 |
| Central Italy | 0.2124 | 0.2042 |
| Southern Italy | 0.3453 | 0.3526 |

| **Head's education** | **SFB** | **SHIW** |
|---|---|---|
| less than 8 years | 0.4173 | 0.4211 |
| compulsory (8 yrs) | 0.2917 | 0.2680 |
| high school | 0.2247 | 0.2381 |
| college degree | 0.0663 | 0.0728 |

| **Head's age** | **SFB** | **SHIW** |
|---|---|---|
| less than 41 | 0.2469 | 0.2114 |
| 41-60 | 0.3967 | 0.4227 |
| 61-80 | 0.3047 | 0.3184 |
| 81+ | 0.0518 | 0.0474 |

consumption (SFB mean and median are 3% and 10% below the corresponding SHIW statistics - the variance is instead lower in SHIW, and so is the overall range). The picture is quite different for non-durable expenditure: mean and median are much (20-25% )higher in SFB compared to SHIW; variability is also higher in SFB, as for food. The comparison looks promising for income, but we know the SFB data is heavily corrected to achieve this result.

There are two key reasons to doubt the SHIW non-durable consumption data. First, is the extreme difficulty of the question. The exact wording is: "What was your family's average monthly expenditure in 1995 for all consumption items? Consider all expenses, including food, but excluding those for: housing maintenance; mortgage installments; purchases of valuables, automobiles, home durables and furniture; housing rent; insurance premiums". This question is then followed by a similar food question ("What was your family's average monthly expenditure for food alone? Consider expenses on all food items in grocery stores or similar food stores and expenses on meals normally consumed out ") and by detailed questions on the other items excluded from non-durable consumption. Second, is the evidence on saving rates computed on the basis of this question. Not only are saving rates unreasonably high (the aggregate saving rate is 23.4%, versus the national accounts equivalent of 16.7% in 1995), but their age profile strongly contradicts what one might expect from theory and evidence for other countries. As shown in Figure 2.1 individual saving rates monotonically increase from 5% to 18% between the ages of 25 and 60, and then stay above 15% for all ages above 60. It is hard to understand why so much saving should take place in old age.

A further problem with the SHIW consumption data appears when we plot histograms for food and non-durable expenditure. In Figures 2.2-2.5 we plot the histograms for food and for non-durable expenditure in SFB and in SHIW: it is apparent that in the SHIW data there is major heaping and rounding taking place at multiples of 500 and 1000. This is a typical problem with recall data, that makes a direct matching of the two surveys difficult.

## 3. The Method

In this paper we assume that the SFB expenditure data is at worst affected by classical measurement error. As for SHIW data, we shall argue that food data is of similar quality, once heaping and rounding are taken into account, while total non-durable expenditure is seriously deficient. We therefore use the SFB information on non-durable expenditure to predict non-durable expenditure in the SHIW sample by adopting a statistical matching technique.

There are a number of methods for statistical matching in the literature (see Baldini, 1998, e.g.) . In order for statistical matching to be feasible we require

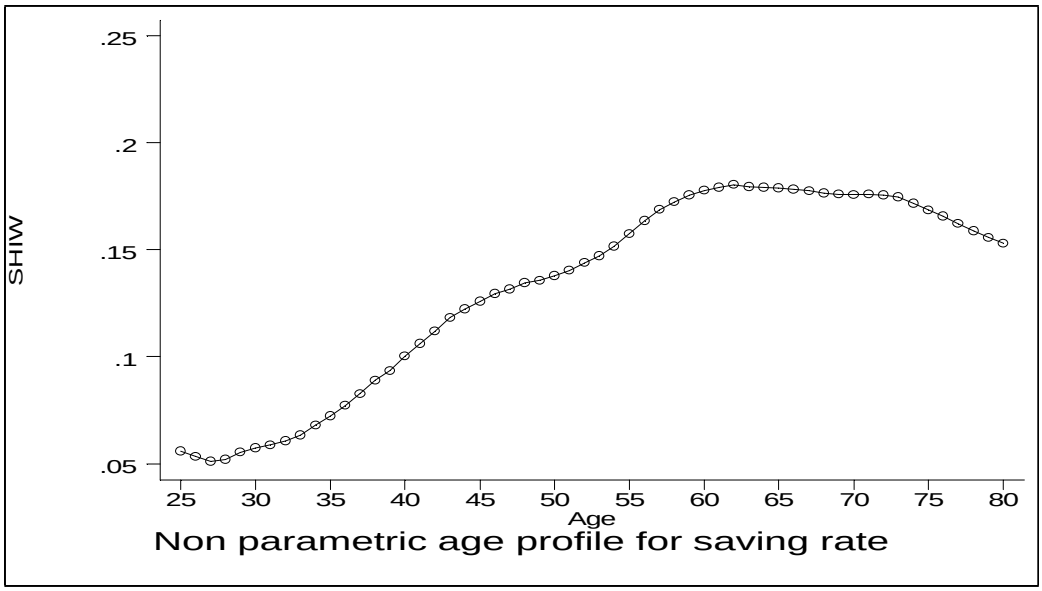


Figure 2.1: Non parametric age profile for saving rates

that:

- surveys should be random samples from the same population
- there is a common set of conditioning variables.

In our case, the first condition is met by design, after allowance is made for sampling and response differences. The second condition is also satisfied (after some recoding, see Rosati (1999): The two surveys share information on household composition, region of residence, age and education of the head, i.e. on valid conditioning variables for the problem under investigation (consumption and savings). It would therefore be possible to use the common reduced form methods suggested by Angrist and Krueger (1992) and Arellano and Meghir (1992).

However, reduced form regressions fail to use information contained in endogenous variables. In our case, we have reliable information on both food and total non-durable expenditure in one of the two surveys, the SFB (and, we shall argue, useful information on food expenditure in the other survey, the SHIW). We therefore estimate a structural relation on SFB data: an inverse Engel curve conditional upon a number of observable household characteristics that are recorded in both surveys. We use information on the quality of the housing stock as additional instruments (that also exist in SHIW). The fact that Engel curve estimation is a well established practice in the economic literature helps us evaluate economically

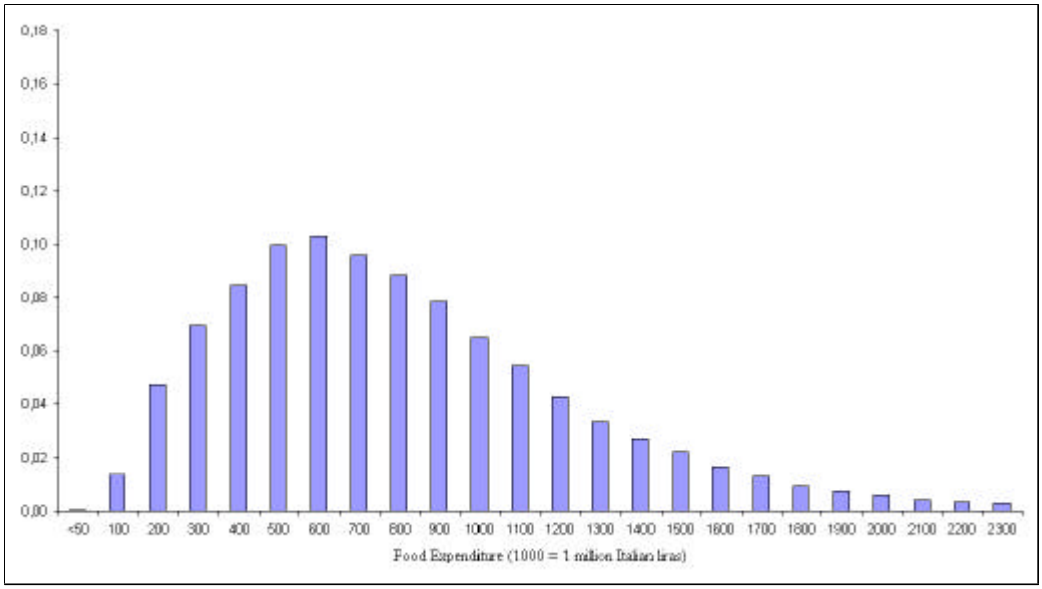


Figure 2.2: Observed food expenditure for SFB data

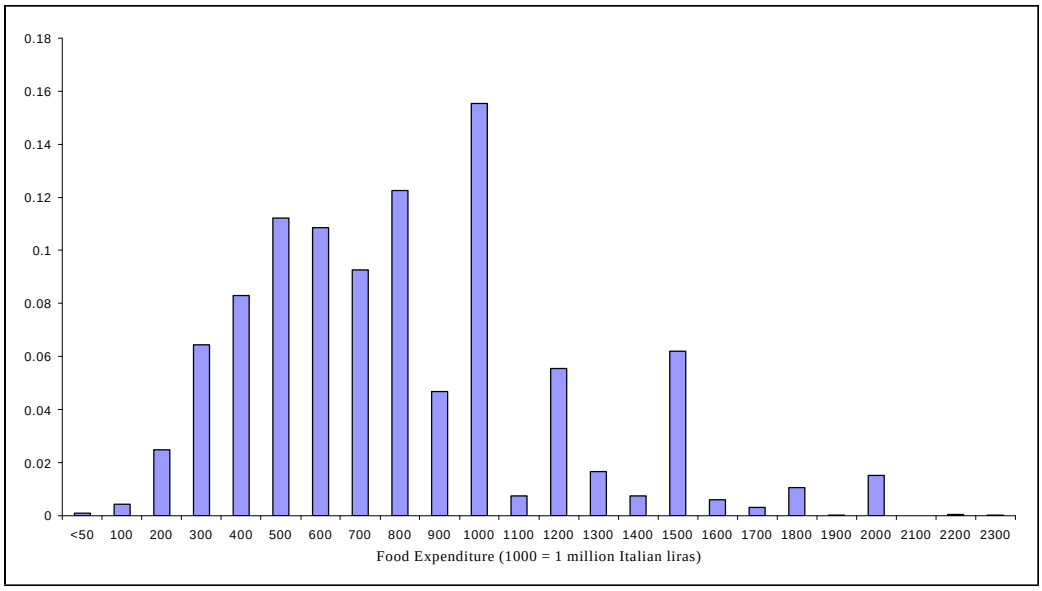


Figure 2.3: Observed food expenditure for SHIW data

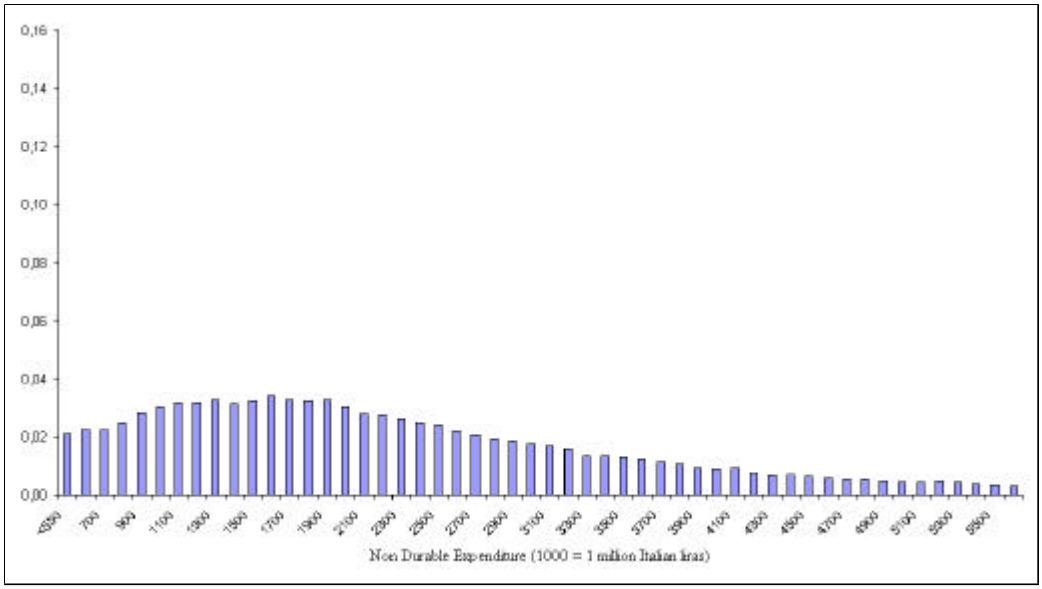


Figure 2.4: Observed non durable expenditure for SFB data

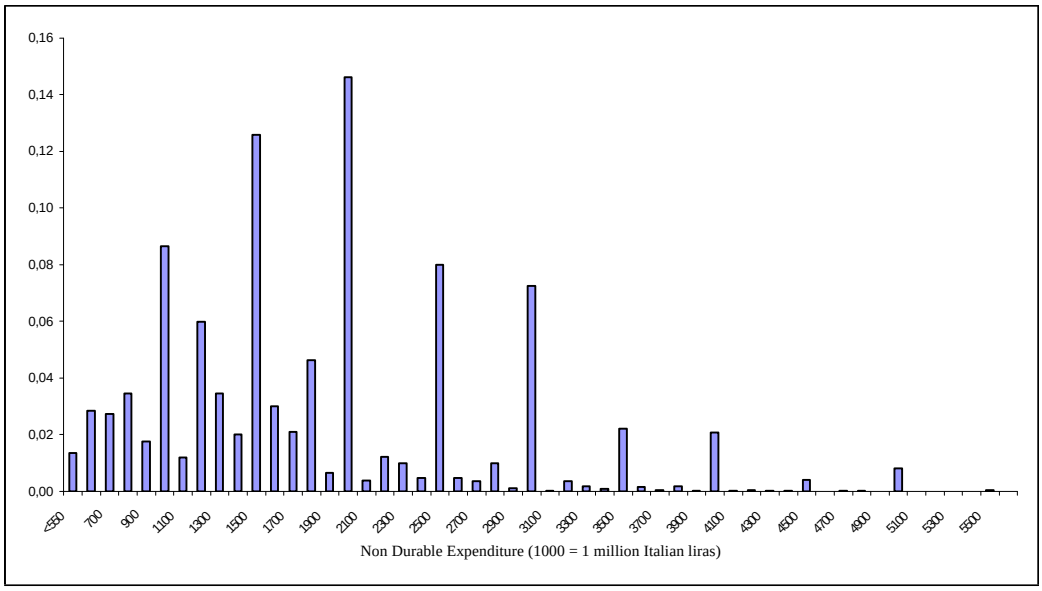


Figure 2.5: Observed non durable expenditure for SHIW data

the success of our matching procedure (see Deaton and Muellbauer (1980), e.g., for further details on Engel curves and their economic interpretation).

Our structural matching exercise is based upon the following system of equations:

$$\begin{aligned} \ln nd &= X\pi_1 + Z\pi_2 + v \\ \ln food &= \alpha \ln nd + X\beta + \varepsilon = \alpha \ln nd + X\beta + \rho v + u \end{aligned} \tag{3.1}$$

where the first equation explains the logarithm of non-durable expenditure as a function of demographic characteristics, $X$, and of wealth indicators, $Z$, in agreement with the standard intertemporal optimization model for consumption. The second equation is an Engel curve, relating food consumption to total non-durable expenditure (the budget) and to demographic characteristics. Simultaneity implies that $\ln nd$ correlates with the equation error, $\varepsilon$. We follow standard practice and decompose the equation error in the reduced form error term for $\ln nd$, $v$, and a residual term, $u$, that is orthogonal to $\ln nd$, $X$ and $v$.

The standard common reduced form approach uses estimates of the first equation based on one sample (SFB in our case) to predict $\ln nd$ in the other (SHIW). We instead exploit the SFB data to jointly estimate the whole system (3.1) and then use SHIW information on $\ln food$, $X$ and $Z$ to predict $\ln nd$ in SHIW. The extra information we use should in principle improve on the statistical quality of the match.

Our method requires inverting the Engel curve, i.e. treating $\ln food$ as an explanatory variable in the specification:

$$\ln nd = \frac{1}{\alpha}[\ln food + X\beta + \varepsilon] \tag{3.2}$$

where allowance must be made for correlation between $\ln food$ and the equation error. To obtain consistent estimates of $\alpha$ and $\beta$ we can take an instrumental variables/2SLS approach and use the $Z$ as additional instruments.

This is equivalent to estimating in the SFB sample a transformation of $\alpha, \beta, \pi_1$ and $\pi_2$ by applying OLS to the equation:

$$\begin{aligned} \ln nd &= \frac{1}{\alpha+\rho}[\ln food + X(\beta - \rho\pi_1) - \rho Z\pi_2 + u] = \\ &\gamma_1 \ln food + X\gamma_2 + Z\gamma_3 + w \end{aligned} \tag{3.3}$$

The final step involves using these parameter estimates to predict $\ln nd$ in SHIW, conditional upon $X$, $Z$ and $\ln food$. However, we have seen that the SHIW observed measure of food is affected by severe rounding and heaping problems. A correction is required before (3.3) can be used to predict $\ln nd$ in SHIW.

### 3.1. Correcting for sampling differences

We saw above that the descriptive statistics of some conditioning variables (such as region, education, age etc.) differ across the two surveys, and that this difference does not disappear when we use sampling weights. This is not surprising because the Bank of Italy survey uses a coarser stratification scheme that does not depend on household size, as discussed in Brugiavini (1996), and because response rates are different across the two surveys. It is therefore necessary to rely on other methods to weigh observations in the two samples, with a view to checking whether the resulting sample density functions of the variables of interest (expenditure, say) are consistent with the hypothesis that they are random draws from the same population.

We account for differences in the composition with respect to some observable characteristics $z$ using the following weighting procedure which builds on Dehejia and Wahba (1999).

Let $Y$ denote expenditure, our variable of interest: let its cumulative marginal distribution in the two survey be denoted by $F_{Y|SFB}$ and $F_{Y|SHIW}$. Differences between $F_{z|SFB}$ and $F_{z|SHIW}$, the cumulative distribution functions of $z$ in the two samples, can be controlled for by choosing as reference population the one described by the Istat (SFB) sample and comparing the expenditure distribution in this population, $F_{Y|SFB}$, to

$$F_{Y|SHIW}^{SFB} = \int F_{Y|z,SHIW}\ dF_{z|SFB} = \int F_{Y|z,SHIW}\ \frac{dF_{z|SFB}}{dF_{z|SHIW}} dF_{z|SHIW},$$

that is the conditional expenditure distribution for Bank of Italy integrated with respect to $F_{z|SFB}$. This expression defines the weighting function:

$$\omega(z) = \frac{dF_{z|SFB}}{dF_{z|SHIW}}$$

whose role is to down-weigh (up-weigh) those households in the Bank of Italy SHIW sample exhibiting characteristics $z$ over represented (under represented) with respect to the reference population (SFB). Applying Bayes theorem, the weights can also be written as

$$\omega\left(z\right) = \frac{e\left(z\right)}{1-e\left(z\right)} \frac{\Pr\left(SHIW\right)}{\Pr\left(SFB\right)},$$

where $e\left(z\right)$ is the *propensity score* as defined by Rosenbaum and Rubin (1983), that is the conditional probability of observing characteristics $z$ in the population represented by the Istat SFB sample. Notice that if the two groups were balanced

with respect to $z$, then the propensity score would not depend on $z$, $\omega(z)$ would be constant over households and the considered distribution function for the Bank of Italy SHIW sample would collapse to the standard one. See Heckman, Ichimura, Smith, Todd (1998) for a review of propensity score based estimators to control for systematic differences with respect to observable characteristics between different groups.

We replace $\omega(z)$ by its sample counterpart assuming a logistic specification for $e(z)$ depending on a large number of demographic indicators and their interactions; $F_{Y|SFB}$ is then estimated by the empirical distribution function while the corresponding estimate for the Bank of Italy SHIW sample is obtained by the ratio

$$\widehat{F}^{SFB}_{Y|SHIW} = \frac{\sum\limits_{i \in SHIW} \widehat{\omega}_i I(y_i \leq y)}{\sum\limits_{i \in SHIW} \widehat{\omega}_i}$$

where $I(A)$ is the index function of event $A$.

### 3.2. Correcting for heaping and rounding errors

In what follows, we assume that the diary-based expenditure data provided in the SFB are at worst affected by classical (zero-mean) measurement error. This identifying assumption allows us to prove that food expenditure can be described by the same parametric distribution across the two data sets, once we account for the stochastic nature of the coarsening process in SHIW data. We also show that the two surveys are of different data quality with respect to total non durable expenditure.

Our estimation procedure can be summarized by the following two steps:

- We at first specify a suitable parametric family depending on the unknown parameter $\vartheta$ both for food and non durable expenditure in the SFB sample. We adopt the Kolmogorov metric as a minimum distance criterium, choosing the distribution that minimizes the uniform norm of the difference between empirical and fitted cumulative density functions within a large class of parametric families.

- We maintain the same parametric specification for (food and non durable) expenditure in the SHIW sample and estimate $\vartheta$ using the maximum likelihood technique suggested by Heitjan and Rubin (1990) to account for heaping and rounding problems.[1]

[1] In this paper we model the unweighted marginal density function for actual expenditures. If sampling differences were found to be non-negligible, it would be necessary to either model the densities conditional upon observable characteristics that are common across the two samples, or to propensity score weigh the marginal density for the SHIW sample.

Following Heitjan and Rubin (1990), assume that the random variable of interest $Y^*$ (expenditure in our case) is distributed according to a density $f(y^*;\vartheta)$ which is a function of the parameter of interest $\vartheta$. If $Y^*$ was available, inference about $\vartheta$ could be drawn directly by standard methods; suppose instead we observe only a subset of the complete data sample space in which the true unobservable data lie. In other words, instead of observing $Y^*$ directly, we only observe a coarse version $Y$ of the variable $Y^*$.

Assume that the degree of coarseness can be summarized by a random variable $G$ whose conditional distribution given $Y^*$ depends on $\gamma$, the parameter of the incompleteness mechanism. This means that the observed variable $Y$ can be expressed as a function of the pair $(Y^*, G)$ or, more formally, that the conditional distribution of $Y$ given the true unobserved data and the value of the coarsening function is degenerate, that is

$$f(y|y^*,g) = \begin{cases} 1 & \text{if } \; y = Y(Y^*,G) \\ 0 & \text{if } \; y \neq Y(Y^*,G) \end{cases} .$$

Let $H(y)$ be the inverse image of $y$ with respect to this application, that is the set of couples $(y^*, g)$ consistent with $y$. In what follows we assume the variable $G$ is not directly observed, but can at best be inferred from the observed value $y$. Therefore, the likelihood function for the parameters $(\vartheta, \gamma)$ in the SHIW sample can be written as

$$\prod_{i \in SHIW} \int f(y_i|y_i^*, g_i) f(g_i|y_i^*;\gamma) f(y_i^*;\vartheta) dg_i dy_i^*,$$

or equivalently

$$\prod_{i \in SHIW} \int_{H(y_i)} f(g_i|y_i^*;\gamma) f(y_i^*;\vartheta) dg_i dy_i^*. \tag{3.4}$$

The relevant parameter $\vartheta$ is estimated for SFB data simply specifying the likelihood function based on $f(y^*;\vartheta)$.

In our analysis we adopt the following parametric specification for $Y^*$

$$f(y^*|\vartheta_1,\vartheta_2,\vartheta_3) = \frac{\vartheta_3 \vartheta_1 (\vartheta_1 y)^{\vartheta_2\vartheta_3 - 1}}{\Gamma(\vartheta_2)} \exp\left\{-(\vartheta_1 y)^{\vartheta_3}\right\}, \tag{3.5}$$

which defines the family of Generalized Gamma distributions[2]. This includes the Weibull distribution ($\vartheta_2 = 1$), the Half Normal distribution ($\vartheta_2 = 1/2$, $\vartheta_3 = 2$),

[2]Using the aforementioned minimum distance criterium, we reject the hypothesis that the cumulative density function has Weibull, Standard Gamma, Lognormal, Inverse Normal and Skewed Normal functional form. This is also borne out by plotting the empirical against the theoretical percentiles.

the ordinary Gamma distributions ($\vartheta_3 = 1$) and, as a limiting special case when $\vartheta_2 \to \infty$, the Lognormal distribution. The associated distribution function can be expressed as

$$F(y|\vartheta_1, \vartheta_2, \vartheta_3) = \Upsilon\left[(\vartheta_1 y)^{\vartheta_3}; \vartheta_2\right]$$

where $\Upsilon(a; \vartheta_2)$ is the incomplete gamma function ratio. See Johnson, Kotz and Balakrishnan (1994) for further details.

We also assume there are three possible types of rounding error: the reported value, $y$, can be the multiple of 1000, 500 or 100 nearest to the true value $y^*$ (in the sequel we shall denote these three types as A1, A2 e A3). It follows that:

$$\begin{aligned} y &\in A1 \Rightarrow |y^* - y| \leq 500 \\ y &\in A2 \Rightarrow |y^* - y| \leq 250 \\ y &\in A3 \Rightarrow |y^* - y| \leq 50 \end{aligned} .$$

As already said, the variable $G$ identifies the type of rounding; assume $G$ to be continuous and let

$$\begin{aligned} g \geq 1 &\Rightarrow y \in A1 \\ 0 \leq g < 1 &\Rightarrow y \in A2 \\ g < 0 &\Rightarrow y \in A3 \end{aligned} ,$$

so that $g \geq 1$ implies rounding to the nearest multiple of 1000, $0 \leq g < 1$ implies rounding to the nearest multiple of 500, and $g < 0$ implies rounding to the nearest multiple of 100.

If we define:

$$\begin{aligned} H_1 &= [y - 500, y + 500) \times [1, +\infty) \\ H_2 &= [y - 250, y + 250) \times [0, 1) \\ H_3 &= [y - 50, y + 50) \times (-\infty, 0) \end{aligned}$$

it follows that

$$H(y) = \begin{cases} H_1 \cup H_2 \cup H_3 & y \in A1 \\ H_2 \cup H_3 & y \in A2 \\ H_3 & y \in A3 \end{cases} .$$

The likelihood function (3.4) is then specified according to the assumptions above, and allows the rounding process to be a function of exogenous observable characteristics, namely age, education and region. It is in fact possible that response care depends on both recall ability and the shadow value of leisure, that will differ across households. Given that $G$ is a continuous variable, the assumed functional form for its conditional distribution is the normal linear regression

$$f(g|y^*; \gamma, \beta) \sim N(\gamma_0 + \gamma_1 y^* + \beta' z; \tau^2), \tag{3.6}$$

where $z$ is the vector of observable characteristics mentioned above. In this sense, this model can be thought as a generalization of the normal selection model proposed in the econometric literature: if $\gamma_1 = 0$ the coarsening mechanism is ignorable, which corresponds to exogenous selection.[3]

## 4. Data analysis.

We report in Table 4.1 estimates of the propensity score function discussed above. The dependent variable takes value 1 if the observation belongs to SFB, 0 otherwise. The explanatory variables include a set of monthly dummies (not reported), household composition indicators, age, employment status and education of the head dummies plus a number of interactions. The key difference across the two surveys is confirmed to lie in the SHIW relative oversampling of households with children aged less than 18, and undersampling of elderly households. However, significant differences are found along several dimensions.

Propensity-score weighted histograms for non-durable expenditure and food expenditure are presented in Figures 4.1 and 4.2 (all expenditure figures are in thousands Italian liras, where Lit 1,000 is approximately \$.5). On the assumption that sampling differences are adequately captured by our propensity score estimates, the remaining differences between the sample distributions of expenditure reflect solely the different nature of measurement error across the two surveys.

From a comparison of 2.3 with 4.1 and 2.5 and 4.2 we draw the conclusion that the propensity score adjustment makes very little difference to the shape of the histograms. This we take as evidence that correcting for sample differences may not be required.

Inspection of Figures 2.3 and 2.5 reveals instead that the SHIW expenditure data suffer from severe heaping and rounding problems. For non-durable expenditure, there are spikes at all multiples of half a million (particularly at Lit 1, 1.5, 2, 2.5, 3 million), even though other spikes are found at Lit 0.8, 1.2 and 1.8 million. For food, there is a spike at Lit 1 million; smaller spikes are also found at Lit. 0.5, 1.5 and 2 million, even though all multiples of 0.1 million are well represented on the left of 0.9 million.

The maximum likelihood estimates for SFB food expenditure are presented in Table 4.2[4] while the corresponding estimates for SHIW are presented in Table

[3] In principle we would like to make the conditional expectation of the $f(.)$ a function of variables likely to affect the recall error process as well as actual expenditure $y^*$. An example could be the time of interview, as suggested in the context of unemployment duration data by Torelli and Trivellato (1993), or other indicators of interview quality.

[4] To make the data comparable across the two surveys, all expenditure data from the Istat SFB have been seasonally adjusted, by taking residuals from regressions on zero-sum monthly

Table 4.1: Propensity score estimates

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| Intercept | 5.0451 | 0.4237 | 11.905 | 0.000 |
| # members 18-26 | -2.6092 | 0.4204 | -6.206 | 0.000 |
| # members 27-40 | -2.6672 | 0.4193 | -6.361 | 0.000 |
| # members 41-60 | -3.0434 | 0.4115 | -7.396 | 0.000 |
| # members 61-70 | -3.5588 | 0.4111 | -8.657 | 0.000 |
| # members over 70 | -4.4512 | 0.4107 | -10.836 | 0.000 |
| Central Italy | -0.7081 | 0.2165 | -3.270 | 0.001 |
| Southern Italy | -0.4600 | 0.1751 | -2.626 | 0.009 |
| Number of children 0-2 | -3.1963 | 0.4061 | -7.871 | 0.000 |
| Number of children 3-5 | -2.0456 | 0.4279 | -4.780 | 0.000 |
| Number of children 6-9 | -3.2903 | 0.3944 | -8.342 | 0.000 |
| Number of children 10-13 | -2.4582 | 0.4068 | -6.042 | 0.000 |
| Number of children 14-17 | -2.2232 | 0.3941 | -5.640 | 0.000 |
| Number of children over 18 | -0.4751 | 0.1481 | -3.207 | 0.001 |
| # retired members | 0.0924 | 0.1056 | 0.875 | 0.382 |
| At least 2 members | -0.2567 | 0.0753 | -3.410 | 0.001 |
| At least 3 members | -0.0295 | 0.0546 | -0.540 | 0.589 |
| At least 4 members | -0.1490 | 0.0454 | -3.282 | 0.001 |
| At least 5 members | -0.0738 | 0.0545 | -1.354 | 0.176 |
| At least 6 members | -0.1053 | 0.1026 | -1.026 | 0.305 |
| At least 7 members | -0.3614 | 0.1730 | -2.089 | 0.037 |
| Sex (male) | -0.0102 | 0.0584 | -0.175 | 0.861 |
| Age of Head $\geq 26$ | 0.0863 | 0.1301 | 0.664 | 0.507 |
| Age of Head $\geq 40$ | 0.1853 | 0.0762 | 2.430 | 0.015 |
| Age of Head $\geq 60$ | 0.3284 | 0.0756 | 4.340 | 0.000 |
| Age of Head $\geq 70$ | 0.6674 | 0.0939 | 7.105 | 0.000 |
| Head Unemployed | -0.6309 | 0.0734 | -8.587 | 0.000 |
| Head Out of the Labor Force | -0.2185 | 0.0436 | -5.010 | 0.000 |
| Education $\geq 8$ | -0.0172 | 0.0514 | -0.335 | 0.737 |
| Education $\geq 13$ | -0.1435 | 0.0533 | -2.689 | 0.007 |
| University Degree | 0.1760 | 0.0820 | 2.146 | 0.032 |
| Total Surface | -0.0059 | 0.0006 | -8.764 | 0.000 |
| Per-capita Surface | -0.0001 | 0.0013 | -0.057 | 0.955 |
| Homeowner | 0.4885 | 0.0291 | 16.743 | 0.000 |
| Owns secondary residence | -1.3874 | 0.0384 | -36.052 | 0.000 |
| Central Italy * #18-26 | 0.5675 | 0.2549 | 2.226 | 0.026 |
| Southern Italy * #18-26 | 0.2487 | 0.2054 | 1.211 | 0.226 |
| Central Italy * #27-40 | 0.6998 | 0.2612 | 2.679 | 0.007 |
| Southern Italy * #27-40 | 0.2213 | 0.2145 | 1.032 | 0.302 |
| Central Italy * #41-60 | 0.6079 | 0.2255 | 2.695 | 0.007 |
| Southern Italy * #41-60 | 0.0243 | 0.1842 | 0.132 | 0.895 |
| Central Italy * #61-70 | 0.9064 | 0.2115 | 4.285 | 0.000 |
| Southern Italy * #61-70 | 0.1454 | 0.1672 | 0.917 | 0.359 |
| Central Italy * #70+ | 0.8148 | 0.2155 | 3.781 | 0.000 |
| Southern Italy * #70+ | 0.4259 | 0.1764 | 2.414 | 0.016 |
| Central Italy * Educ. $\geq 8$ | 0.2372 | 0.0914 | 2.596 | 0.009 |
| Southern Italy * Educ $\geq 8$ | 0.1673 | 0.0769 | 2.175 | 0.030 |
| Central Italy * Educ $\geq 13$ | -0.0218 | 0.0953 | -0.229 | 0.819 |
| Southern Italy * Educ $\geq 13$ | 0.2047 | 0.0829 | 2.469 | 0.014 |
| Central Italy * Un. degree | 0.2212 | 0.1502 | 1.473 | 0.141 |
| Southern Italy * Un. degree | -0.1101 | 0.1273 | -0.865 | 0.387 |
| Central Italy * Sex | -0.0255 | 0.0911 | -0.280 | 0.779 |
| Southern Italy * Sex | 0.1740 | 0.0778 | 2.235 | 0.025 |

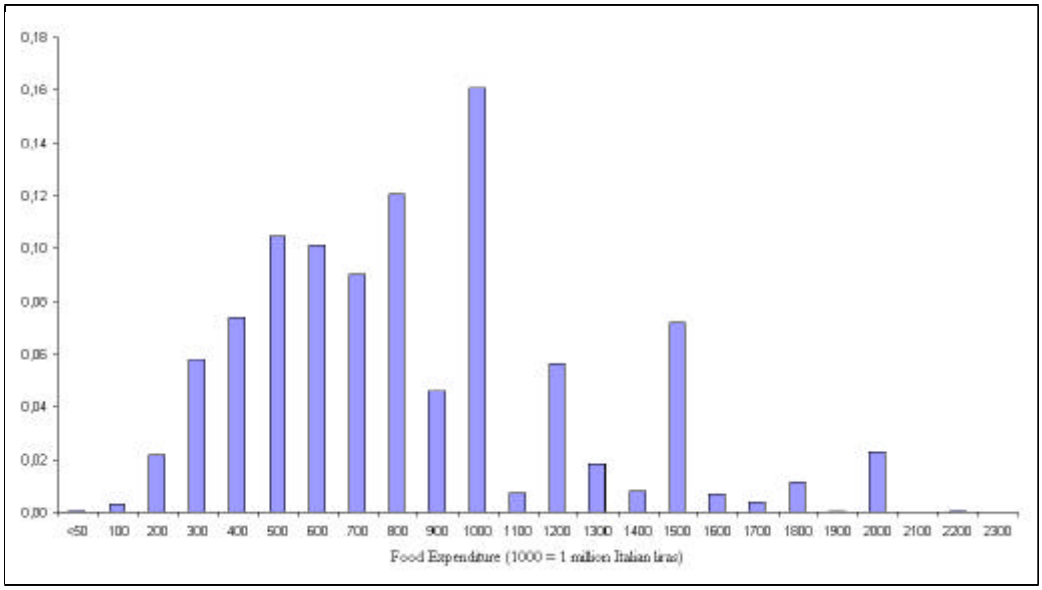


Figure 4.1: Observed (propensity score weighted) food expenditure for SHIW data

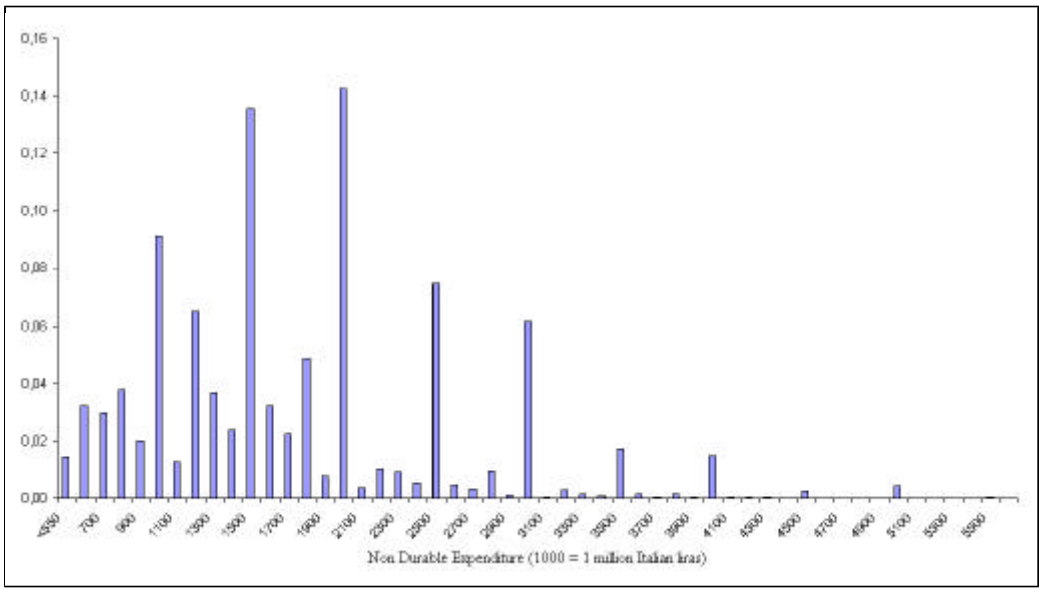


Figure 4.2: Observed (propensity score weighted) non durable expenditure for SHIW data

Table 4.2: SFB Food Data: Density Function Parameters

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| $\vartheta_1^{SFB}$ | 0.0033 | 0.0002 | 16.089 | 0.0000 |
| $\vartheta_2^{SFB}$ | 2.8118 | 0.1134 | 24.791 | 0.0000 |
| $\vartheta_3^{SFB}$ | 1.0877 | 0.0244 | 44.568 | 0.0000 |

Table 4.3: SHIW Food Data: Density Function Parameters

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| $\vartheta_1^{SHIW}$ | 0.0030 | 0.0004 | 7.115 | 0.0000 |
| $\vartheta_2^{SHIW}$ | 3.0614 | 0.3005 | 10.189 | 0.0000 |
| $\vartheta_3^{SHIW}$ | 1.1881 | 0.0636 | 18.694 | 0.0000 |
| $\gamma_0$ | -2.5721 | 0.0996 | -25.822 | 0.0000 |
| $\gamma_1$ | 0.0016 | 0.0001 | 23.896 | 0.0000 |
| At least 8 years | 0.0563 | 0.0685 | 0.823 | 0.2053 |
| At least high school | 0.0574 | 0.0674 | 0.850 | 0.1976 |
| College Degree | 0.1706 | 0.0861 | 1.981 | 0.0238 |
| Central Italy | 0.1135 | 0.0634 | 1.789 | 0.0368 |
| Southern Italy | 0.0966 | 0.0560 | 1.726 | 0.0422 |
| Age 40-60 | 0.1304 | 0.0754 | 1.729 | 0.0419 |
| Age over 60 | 0.1582 | 0.0830 | 1.977 | 0.0240 |

4.3; the table reports estimates of the parameters of the heaping function as well.

The adopted specification for the heaping function (3.6) allows us to establish that the stochastic nature of the coarsening mechanism cannot be ignored in drawing inferences about the parameter of interest $\vartheta$. Maximum likelihood estimates of the parameter $\gamma$ support the idea that the reported expenditure is not *coarsened at random*[5]; a higher expenditure level increases the probability of large rounding errors ($\gamma_1$ is positive and significantly different from zero) and respondents with a college degree or beyond retirement age are also more likely to round off numbers.

A formal test of parameter equality across the two samples fails to reject the

dummies.

[5]If observations were coarsened at random, the likelihood inference for $\vartheta$ would be drawn treating the reported expenditures as if they were simple grouped data. See Heitjan and Rubin (1991) on how to extend the notion of missing at random to more complicated incomplete data problems.

Table 4.4: SFB Non-durable Data: Density Function Parameters

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| $\vartheta_1^{SFB}$ | 0.6233 | 0.1540 | 4.0681 | 0.0000 |
| $\vartheta_2^{SFB}$ | 15.9711 | 0.6180 | 25.842 | 0.0000 |
| $\vartheta_3^{SFB}$ | 0.3839 | 0.0077 | 50.099 | 0.0000 |

Table 4.5: SHIW Non-durable Data: Density Function Parameters

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| $\vartheta_1^{SHIW}$ | 0.0161 | 0.0074 | 2.162 | 0.0153 |
| $\vartheta_2^{SHIW}$ | 9.0446 | 1.2874 | 7.0260 | 0.0000 |
| $\vartheta_3^{SHIW}$ | 0.6597 | 0.0476 | 13.854 | 0.0000 |
| $\gamma_0$ | -1.2799 | 0.0705 | -18.167 | 0.0000 |
| $\gamma_1$ | 0.0006 | 0.0001 | 26.670 | 0.0000 |
| At least 8 years | 0.0162 | 0.0499 | 0.3241 | 0.3729 |
| At least high school | 0.1456 | 0.0505 | 2.882 | 0.0020 |
| College Degree | 0.0627 | 0.0728 | 0.862 | 0.1943 |
| Central Italy | -0.0128 | 0.0487 | -0.262 | 0.3967 |
| South Italy | 0.0577 | 0.0436 | 1.322 | 0.0930 |
| Age 40-60 | 0.0438 | 0.0515 | 0.8512 | 0.1974 |
| Age over 60 | -0.0359 | 0.0569 | -0.630 | 0.2643 |

null: the ML test statistic of joint parameter equality ($H_0 : \vartheta^{SFB} = \vartheta^{SHIW}$) takes a value of 4.47 that compares to a critical value of 7.82 ($= \chi_3^2(0.95)$).

When we perform a similar exercise on total non-durable expenditure, we still find that a generalized Gamma provides an adequate fit. The actual parameter estimates for SFB and SHIW are reported in Tables 4.4 and 4.5 respectively. It's worth noting that the parameter estimates of the coarsening function are poorly determined in this case. A formal test of the gamma function parameter equality across the two samples strongly rejects the null.

On the basis of the above, we conclude that the food data are of comparable quality and information content across the two surveys, once heaping and rounding are accounted for. A different conclusion must be drawn for non-durable expenditure.

We create multiple imputations of food and non-durable expenditure for the Bank of Italy SHIW sample implementing an acceptance-rejection procedure based

on the fitted model. Since by applying Bayes theorem we have:

$$f(y^*, g|y; \vartheta, \gamma) \propto \begin{cases} f(y^*, g; \vartheta, \gamma) & \text{if } \; y = Y(y^*, g) \\ 0 & \text{if } \; y \neq Y(y^*, g) \end{cases} ,$$

for each unit we draw a couple $(y^*, g)$ from the estimated distribution $f(y^*, g; \widehat{\vartheta}, \widehat{\gamma})$ until $(y^*, g) \in H(y)$, that is until the generated couple is consistent with the observed value $y$. We then impute $y^*$ as the true value of the observed expenditure $y$.

The average histograms of 100 imputations are shown in Figures 4.3 and 4.4 and can be compared to the corresponding distributions of food and non-durable expenditure for the Istat SFB sample. Different compositions respect to observable characteristics are corrected by propensity score weighting the imputed observations to obtain the same distribution of observable characteristics for the two samples.

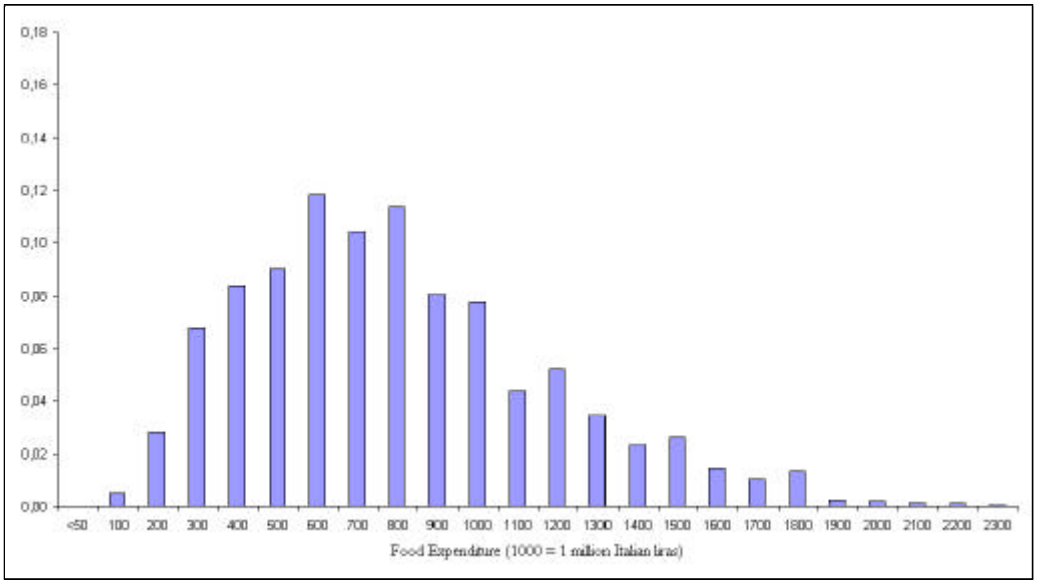


Figure 4.3: Imputed (propensity score weighted) food expenditure for SHIW data.

## 5. Estimates of the inverse Engel Curve

In this section we present non-parametric, semi-parametric and parametric estimates for the inverse Engel curve (3.2) using SFB data. Even though in prediction

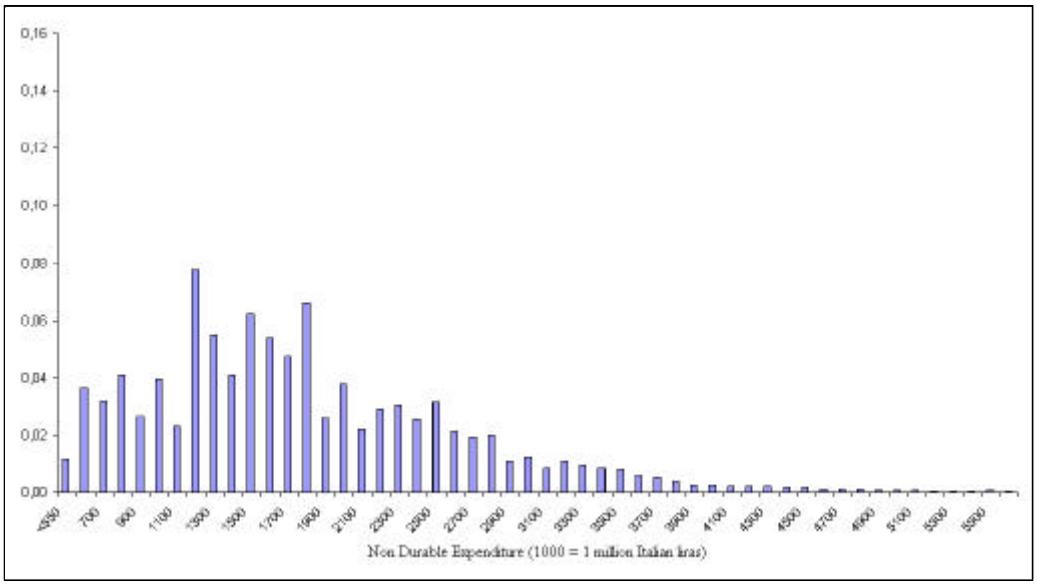


Figure 4.4: Imputed (propensity score weighted) non durable expenditure for SHIW data.

we use (3.3), specification (3.2) is easier to interpret and to relate to economic theory and standard econometric practice.

We first show that a non-parametric version of (3.2) that fails to condition upon demographics is close to a straight line, and its slope is everywhere less than one. Conditioning upon a set of demographic characteristics is more easily done in a semiparametric context. In that context we can also tackle issue of potential simultaneity bias, by instrumenting the $ln(food)$ term consistently with (3.1). The semi-parametric OLS-equivalent is close to linear, and its slope is less than unity. The semi-parametric IV estimates are very close to a straight line with slope in the 1.5 region.

On the basis of these results, we go on to estimate parametrically a double log specification for (3.2). This gives us an idea of the fit of the equation and of the quality of the chosen instruments. On both counts we find that the specification is quite good, and thus suitable for predicting on the SHIW sample.

### 5.1. Non-parametric estimates

We present non-parametric estimates for both the curve (in double log form) and its derivative (i.e.: the inverse elasticity), obtained by applying to the SFB data the local polynomial fitting techniques described in Fan and Gijbels (1996).

Let $m(x)$ be the regression function for (log-) non durable expenditure given

the value $x$ of (log-) food expenditure and let $m^{(j)}(x)$ its $j^{th}$-derivative. Even if its functional form is unknown, $m(x)$ can always be locally approximated by a polynomial of suitable degree using a Taylor series expansion. Assume that the $(p+1)^{th}$ derivative of $m(x)$ at $x_0$, say $m^{(p+1)}(x_0)$, exists; then, for $x$ in a neighborhood of $x_0$,

$$m(x) \sim \sum_{j=0}^{p} \frac{m^{(j)}(x_0)}{j!}(x - x_0)^j = \sum_{j=0}^{p} \beta_j (x - x_0)^j. \tag{5.1}$$

This polynomial is then fitted locally as a weighted least squares regression problem minimizing the loss function

$$\sum_i \left\{ \ln nd_i - \sum_{j=0}^{p} \beta_j (\ln food_i - x_0)^j \right\}^2 K(\frac{\ln food_i - x_0}{h}),$$

where $K(.)$ denotes a kernel function assigning a weight to each observation and $h$ is a bandwidth.

If $h$ is small, the local linear fitting process depends heavily on those observations that are closest to $x_0$ and tends to give a less smooth estimate; in this sense, as $h$ becomes closer to zero the estimator tends towards interpolation of the data. On the other hand, a larger $h$ tends to weigh the observations more equally and as $h$ increases the estimate tends towards the ordinary least squares line through the data.

Expression (5.1) suggests that an estimator for $m^{(j)}(x_0)$, $j \geq 0$, is

$$\widehat{m}^{(j)}(x_0) = j!\widehat{\beta}_j;$$

the whole curve is then obtained running the above described procedure with $x_0$ varying in an appropriate domain for (log-)food expenditure in the SFB sample.

In what follows we present results from a local polynomial fitting of order three for the SFB regression function and its derivative using the Epanechnikov kernel $K(x) = 0.75(1 - x^2)_+$ and different values of the bandwidth $h$[6].

Figure 5.1 shows that the inverse Engel curve is close to linear over the chosen range (that is defined to include all points between the first and the 99th percentile), while Figure 5.2 reveals that the average derivative lies mostly in the .5-.98 interval[7]. This latter finding would imply that food is a luxury if we could

[6]It can be shown that this kernel is optimal in the sense that it minimizes the asymptotic mean squared error of the resulting polynomial estimator.

[7]Ideal theoretical choices for the $h$ parameter are easy to obtain even if they are not always readily usable, since they depend on unknown quantities. Our conclusions are robust with respect to a wide range of different values of $h$. We therefore present graphs relative to an optimal bandwidth obtained by a cross validation method, as suggested in the literature.

disregard the less than perfect fit of the regression function. We shall argue in the sequel that this finding is also attributable to simultaneity bias.[8]

### 5.2. Semi-parametric estimates

The non-parametric evidence obtained suggests that the statistical relation between $\ln nd$ and $\ln food$ may be close to linear. However, what interests us is the relation conditional upon demographic characteristics as in (3.2). Further, we want to take into account the likely correlation between the explanatory variable $\ln food$ and the error term.

Non-parametric analysis conditional upon a number of variables is notoriously difficult (this is known as the curse of dimensionality). A simple way out is to resort to semi-parametric estimation instead. Also, in a semiparametric context the instrumental variables estimator can be easily implemented, as discussed in Blundell, Duncan and Pendakur (1998) (see also Newey, Powell and Vella 1999).

We consider a small set of demographic variables (a 3rd-degree polynomial in #household members; a 2nd degree polynomial in age of the head; a set of ratios of #household members within age range to total #household members) that enter linearly in the inverse Engel curve:

$$\ln nd_i = \beta' x_i + g(\ln food_i) + \zeta_i$$

We allow for the potential correlation between $ln(food)$ and the error term by using an augmented regression technique. We assume the following linear conditional model:

$$\ln(food_i) = \delta' x_i + \varphi z_i + \xi_i$$

where the $z$ are additional instruments (we use total and per-capita housing surface as instruments). We know that in the fully linear model these are valid instruments, in the sense that the oderidentifying restrictions implied by them are not rejected. We add linearly to the first equation the estimated residual,$\hat{\xi}_i$, and its square, and estimate the resulting augmented regression by a local polynomial of order 1.

[8] This analysis does not condition upon observable characteristics, as fully non-parametric estimation suffers from the curse of dimensionality. We therefore also estimated semi-parametrically the regression function, conditioning upon age, family size and household composition indicators. In this context we could also implement an Instrumental Variables estimation technique, where the key instruments are housing wealth indicators (see Blundell Duncan and Pendakur, 1998). For both OLS and IV the regression function is almost perfectly linear. In the former case the slope is less than 1, in the latter case is around 2.

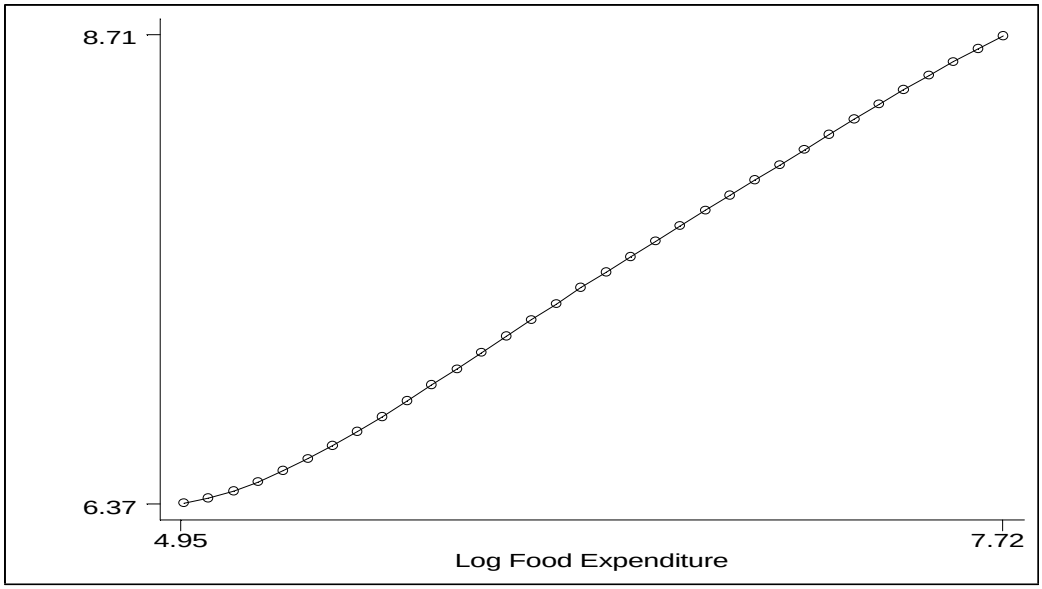


Figure 5.1: Non parametric estimate of inverse Engel Curve

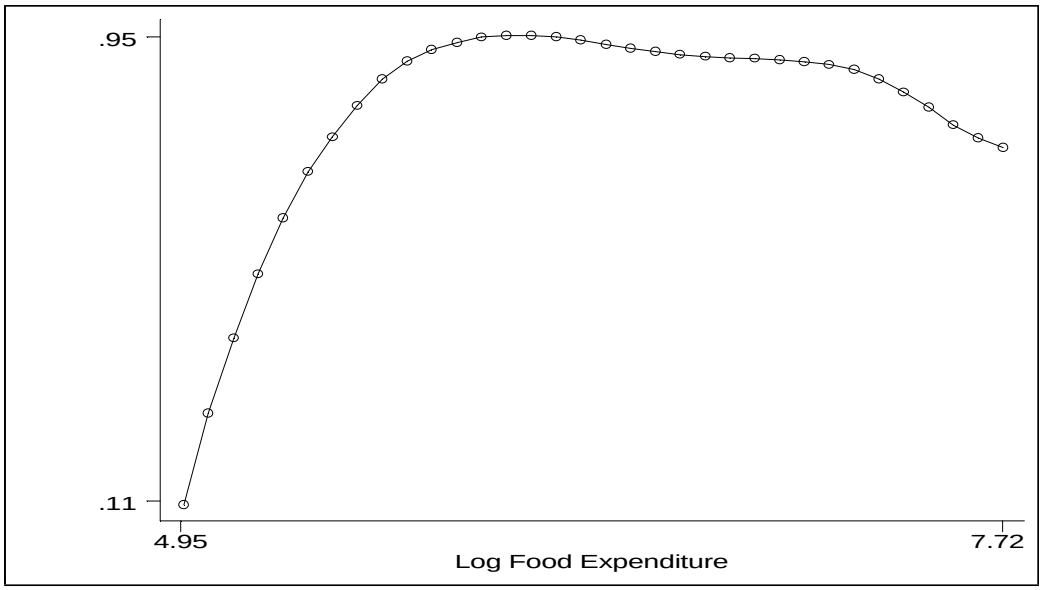


Figure 5.2: Non parametric estimate of inverse elasticity

Estimation results are shown in Figures 5.3 and 5.4: g(.) is close to linearity both when we don't and we do allow for simultaneity. The slope of the inverse Engel curve is less than one in the former case, it is larger than one in the latter case.

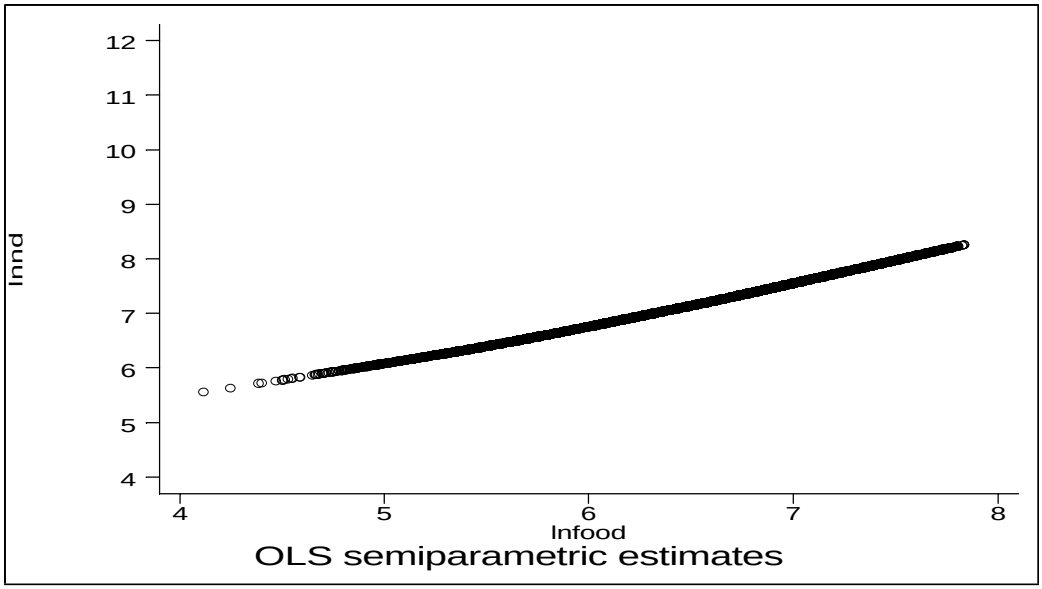


Figure 5.3: Estimated Inverse Engel Curve

### 5.3. Parametric estimates

In Table 5.1 we present OLS estimates of a double-log specification. As we have seen above, the non-parametric and semi-parametric estimates of the inverse Engel curve support the view that a double-log linear specification is adequate. Given that our goal is prediction parametric estimation is preferable. It also allows us to control for a large number of social-demographic indicators, thus gaining precision.

The demographic indicators we use as controls are: region of residence, household composition indicators, education, sex and age of the head, and their interactions. A spline function of age is also interacted with the $\ln(food)$ variable, to allow for an age-dependent elasticity (the cutoff points for the age groups are 27, 41, 61 and 71). In Table 5.1 we present OLS estimates of the inverse Engel curve. The fit of the equation is good (63.49% of the variance is explained by the model), and the key parameter (the coefficient on $\ln(food)$) is precisely estimated at .707. If there was a perfect fit, this would imply that for households whose head

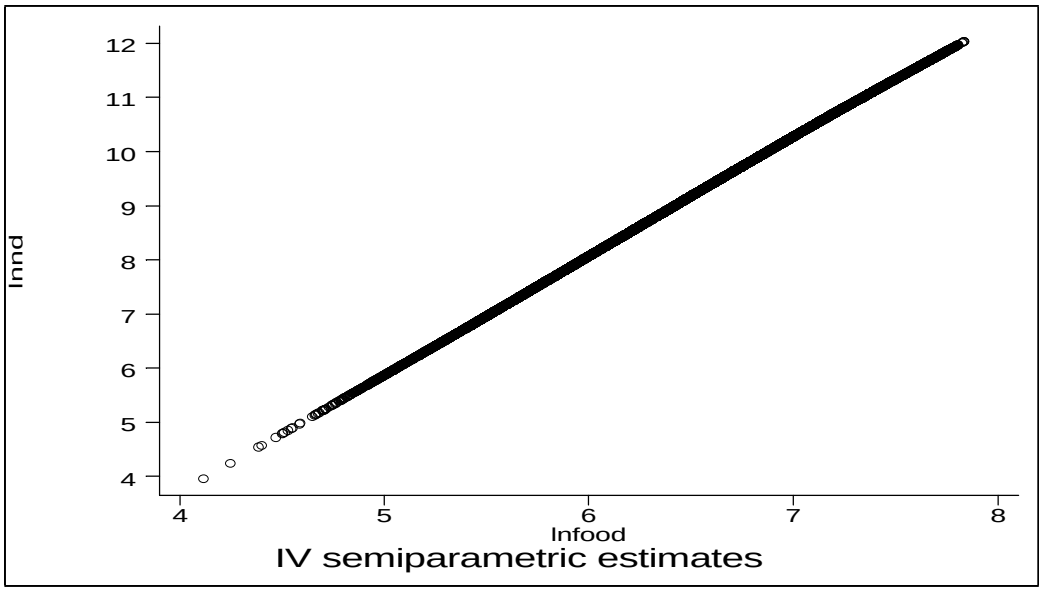


Figure 5.4: Estimated Inverse Engel Curve

is less than 27 years of age the elasticity of food expenditure to total non-durable expenditure is 1.41, an implausibly high number. Even if we consider that the $R^2$ is less than one, we obtain an implied elasticity in the .9 region, much higher than normally found. Similar inference can be drawn for other age groups (the reciprocals of the point estimates for households in age group 1, ranging 27-40, is 1.52; for age group 2, ranging 41-60, 1.38; and for age groups 3 and 4 - 61-70 and beyond - is 1.29)[9].

This counter-intuitive evidence may be due to simultaneity problems: first, food expenditure is a component of total non-durable expenditure; secondly, food and other non-durable expenditure are jointly determined and reflect the overall standard of living for each household (they both depend, via the marginal utility of wealth, on total lifetime resources available to the household).

We therefore estimate the inverse Engel curve by Instrumental Variables: we treat ln(Food Expenditure) and its interactions with age-group dummies as endogenous and we use as additional instruments two variables that capture the quality of housing available to the consumers (these are the total surface of the house and the per-capita surface) as well as their interaction with the same age-group dummies. The idea is that these variables correlate with the long-term standard of living the household can afford, i.e. that they belong in the first

[9]If we consider a specification without age interactions, the $R^2$ is 63.39% and the parameter of interest is estimated to be .725 (s.e.: .005). Estimation results are available upon request.

Table 5.1: Inverse Engel Curve Estimates

| | | | | | |
|---|---|---|---|---|---|
| Dependent variable | $\ln(nd)$ | Observations | 33205 | F(64,33140) | 900.36 |
| Estimation Method | OLS | $R^2$ | 0.6349 | Root MSE | 0.3949 |

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| Intercept | 2.8058 | 0.2013 | 13.932 | 0.000 |
| $\ln(food)$ | 0.7075 | 0.0305 | 23.179 | 0.000 |
| $\ln(food)$* Age $\geq 26$ | -0.0496 | 0.0318 | - 1.559 | 0.119 |
| $\ln(food)$* Age $\geq 40$ | 0.0632 | 0.0121 | 5.207 | 0.000 |
| $\ln(food)$* Age $\geq 60$ | 0.0443 | 0.0123 | 3.594 | 0.000 |
| $\ln(food)$* Age $\geq 70$ | 0.0101 | 0.0142 | 0.707 | 0.480 |
| # members 18-26 | 0.1207 | 0.0505 | 2.387 | 0.017 |
| # members 27-40 | 0.0794 | 0.0500 | 1.589 | 0.112 |
| # members 41-60 | 0.0392 | 0.0491 | 0.799 | 0.425 |
| # members 61-70 | -0.0386 | 0.0500 | -0.774 | 0.439 |
| # members over 70 | -0.1410 | 0.0508 | -2.776 | 0.006 |
| Central Italy | -0.0272 | 0.0357 | -0.762 | 0.446 |
| Southern Italy | -0.1771 | 0.0291 | -6.080 | 0.000 |
| Number of children 0-2 | -0.1341 | 0.0499 | -2.685 | 0.007 |
| Number of children 3-5 | -0.1091 | 0.0526 | -2.076 | 0.038 |
| Number of children 6-9 | -0.0951 | 0.0479 | -1.985 | 0.047 |
| Number of children 10-13 | -0.0238 | 0.0499 | -0.477 | 0.633 |
| Number of children 14-17 | 0.0162 | 0.0472 | 0.345 | 0.730 |
| Number of children over 18 | 0.0589 | 0.0238 | 2.469 | 0.014 |
| # retired members | 0.0395 | 0.0187 | 2.107 | 0.035 |
| At least 2 members | 0.0670 | 0.0087 | 7.702 | 0.000 |
| At least 3 members | 0.0601 | 0.0083 | 7.171 | 0.000 |
| At least 4 members | 0.0287 | 0.0079 | 3.894 | 0.000 |
| At least 5 members | -0.0039 | 0.0095 | -0.408 | 0.683 |
| At least 6 members | -0.0482 | 0.0187 | -2.575 | 0.010 |
| At least 7 members | 0.0343 | 0.0360 | 0.952 | 0.341 |
| Sex (male) | -0.0062 | 0.0095 | -0.654 | 0.513 |
| Age of Head $\geq 26$ | 0.3700 | 0.2063 | 1.793 | 0.073 |
| Age of Head $\geq 40$ | -0.3936 | 0.0835 | -4.714 | 0.000 |
| Age of Head $\geq 60$ | -0.2906 | 0.0854 | -3.402 | 0.001 |
| Age of Head $\geq 70$ | -0.0336 | 0.0971 | -0.346 | 0.729 |
| Head Unemployed | -0.1377 | 0.0148 | -9.281 | 0.000 |
| Head Out of the Labor Force | -0.0383 | 0.0076 | -5.026 | 0.000 |
| Education $\geq 8$ | 0.0804 | 0.0086 | 9.267 | 0.000 |
| Education $\geq 13$ | 0.0830 | 0.0090 | 9.172 | 0.000 |
| University Degree | 0.1008 | 0.0144 | 6.978 | 0.000 |
| Central Italy * #18-16 | 0.0141 | 0.0424 | 0.334 | 0.738 |
| Southern Italy * #18-16 | -0.0559 | 0.0343 | -1.629 | 0.103 |
| Central Italy * #27-40 | -0.0130 | 0.0411 | -0.317 | 0.751 |
| Southern Italy * #27-40 | -0.0130 | 0.0411 | -0.317 | 0.751 |
| Central Italy * #41-60 | -0.0417 | 0.0370 | -1.126 | 0.260 |
| Southern Italy * #41-60 | -0.0145 | 0.0306 | -0.473 | 0.636 |
| Central Italy * #61-70 | -0.0056 | 0.0351 | -0.160 | 0.873 |
| Southern Italy * #61-70 | 0.0097 | 0.0289 | 0.338 | 0.735 |
| Central Italy * #70+ | -0.0096 | 0.0362 | -0.267 | 0.789 |
| Southern Italy * #70+ | 0.0116 | 0.0298 | 0.392 | 0.695 |
| Central Italy * Educ. $\geq 8$ | -0.0239 | 0.0151 | -1.579 | 0.114 |
| Southern Italy * Educ $\geq 8$ | 0.0057 | 0.0131 | 0.437 | 0.662 |
| Central Italy * Educ $\geq 13$ | 0.0198 | 0.0160 | 1.241 | 0.215 |
| Southern Italy * Educ $\geq 13$ | 0.0346 | 0.0139 | 2.496 | 0.013 |
| Central Italy * Un. degree | 0.0131 | 0.0250 | 0.523 | 0.601 |
| Southern Italy * Un. degree | 0.0189 | 0.0220 | 0.856 | 0.392 |
| Central Italy * Sex | 0.0164 | 0.0146 | 1.124 | 0.261 |

equation in (3.1).

In Table 5.2 we present estimation results. Even though the estimated standard errors of the estimates are larger than those shown in Table 5.1, inference can still be drawn with good confidence. The estimated elasticity of food is 0.40 for the youngest, 0.41 for group 1, 0.46 for group 2, 0.48 for groups 3 and 4. These estimates are fully consistent with the notion that food is a necessity[10].

Standard goodness of fit measures do not apply in the Instrumental Variables context. However, IV estimates can be obtained by the two stage least squares procedure: the endogenous regressors are replaced by their fitted values from regressions on the full instruments set. The equation is then estimated by OLS, and its coefficient of determination is the generalized $R^2$ ($GR2$) statistic of Pesaran and Smith (1994). In our case $GR2$ is 41.53%, suggesting that the overall equation fit is quite good[11].

A formal test of instruments validity fails to reject the null (the Sargan criterion is 7.71 and is distributed as a chi-squared statistic with 5 degrees of freedom under the null of instruments validity). A Hausman test strongly rejects the null of equality of OLS and IV coefficients, while an F test also rejects the null that the same $\ln(food)$ coefficient applies to all age groups (its p-value is .0036).

## 6. Predictions

We have argued above that the Istat SFB data set contains reliable information on expenditure items. In particular, we have constructed food and total non-durable expenditure aggregates that are diary-based and are defined in a way that is fully comparable to the Bank of Italy SHIW corresponding items. Also, we have shown that the type of rounding and heaping errors typical of recall questions can be dealt with in estimation, and that the underlying density function is statistically the same for food expenditure, but differs markedly for total non-durable expenditure. Finally, we have seen that an inverse Engel curve can be successfully estimated on the SFB data, and the key estimated parameters are in line with what is normally found in other diary-based household data sets.

[10] A reassuring feature of this set of estimates is that the estimated direct Engel curve also implies elasticities for food of approximately 0.4-0.5 . This can be taken as evidence that our sample is sufficiently large for us to rely on asymptotic properties of the estimator (IV is not invariant to normalization in finite samples).

When we drop the age group interaction terms from both the set of explanatory variables and the set of instrumental variables, we obtain a point estimate of the parameter of interest of 2.00 (s.e.: 0.098), implying a budget elasticity for food of 0.50 .

[11] The relevant measure of goodness of fit for this specification is $R^2$ obtained when estimating by OLS equation (3.3). This turns out to be 64.12%.

Table 5.2: Inverse Engel Curve Estimates

| | | | | | |
|---|---|---|---|---|---|
| Dependent variable | $\ln(nd)$ | Observations | 33205 | F(64,33205) | 161.45 |
| Estimation Method | IV | Generalized $R^2$ | .4153 | Root MSE | 0.7542 |

| **Parameters** | **Estimates** | **Std. Errors** | **t-values** | **Prob.** |
|---|---|---|---|---|
| Intercept | -7.9507 | 1.3085 | -6.076 | 0.000 |
| $\ln(food)$ | 2.4789 | 0.2085 | 11.888 | 0.000 |
| $\ln(food)$* Age $\geq 26$ | -0.0792 | 0.1678 | -0.472 | 0.637 |
| $\ln(food)$* Age $\geq 40$ | -0.2461 | 0.0810 | -3.036 | 0.002 |
| $\ln(food)$* Age $\geq 60$ | -0.0631 | 0.0592 | -1.066 | 0.286 |
| $\ln(food)$* Age $\geq 70$ | -0.0072 | 0.0645 | -0.113 | 0.910 |
| # members 18-26 | 0.0642 | 0.0978 | 0.657 | 0.511 |
| # members 27-40 | -0.1279 | 0.0982 | -1.302 | 0.193 |
| # members 41-60 | -0.2140 | 0.0967 | -2.213 | 0.027 |
| # members 61-70 | -0.1646 | 0.0985 | -1.671 | 0.095 |
| # members over 70 | -0.0968 | 0.1020 | -0.949 | 0.343 |
| Central Italy | -0.0616 | 0.0684 | -0.901 | 0.368 |
| Southern Italy | -0.1512 | 0.0557 | -2.713 | 0.007 |
| Number of children 0-2 | 0.2349 | 0.1003 | 2.342 | 0.019 |
| Number of children 3-5 | -0.0018 | 0.1014 | -0.018 | 0.986 |
| Number of children 6-9 | 0.1085 | 0.0933 | 1.163 | 0.245 |
| Number of children 10-13 | 0.1579 | 0.0967 | 1.632 | 0.103 |
| Number of children 14-17 | 0.0537 | 0.0905 | 0.594 | 0.553 |
| Number of children over 18 | -0.0841 | 0.0474 | -1.775 | 0.076 |
| # retired members | -0.0232 | 0.0379 | -0.613 | 0.540 |
| At least 2 members | -0.4080 | 0.0423 | -9.633 | 0.000 |
| At least 3 members | -0.2759 | 0.0317 | -8.686 | 0.000 |
| At least 4 members | -0.1946 | 0.0230 | -8.448 | 0.000 |
| At least 5 members | -0.0952 | 0.0197 | -4.819 | 0.000 |
| At least 6 members | -0.1650 | 0.0370 | -4.452 | 0.000 |
| At least 7 members | -0.0607 | 0.0692 | -0.877 | 0.381 |
| Sex (male) | -0.1576 | 0.0221 | -7.118 | 0.000 |
| Age of Head $\geq 26$ | 0.5547 | 1.0831 | 0.512 | 0.609 |
| Age of Head $\geq 40$ | 1.6143 | 0.5486 | 2.942 | 0.003 |
| Age of Head $\geq 60$ | 0.3999 | 0.4054 | 0.986 | 0.324 |
| Age of Head $\geq 70$ | 0.0571 | 0.4337 | 0.132 | 0.895 |
| Head Unemployed | 0.2017 | 0.0401 | 5.026 | 0.000 |
| Head Out of the Labor Force | 0.0626 | 0.0168 | 3.723 | 0.000 |
| Education $\geq 8$ | -0.0031 | 0.0179 | -0.175 | 0.861 |
| Education $\geq 13$ | -0.0046 | 0.0187 | -0.249 | 0.804 |
| University Degree | 0.0362 | 0.0281 | 1.288 | 0.198 |
| Central Italy * #18-16 | 0.1980 | 0.0828 | 2.391 | 0.017 |
| Southern Italy * #18-16 | 0.1850 | 0.0685 | 2.700 | 0.007 |
| Central Italy * #27-40 | 0.1367 | 0.0796 | 1.717 | 0.086 |
| Southern Italy * #27-40 | 0.1741 | 0.0682 | 2.552 | 0.011 |
| Central Italy * #41-60 | 0.0819 | 0.0715 | 1.146 | 0.252 |
| Southern Italy * #41-60 | 0.2431 | 0.0622 | 3.903 | 0.000 |
| Central Italy * #61-70 | 0.0850 | 0.0675 | 1.259 | 0.208 |
| Southern Italy * #61-70 | 0.2599 | 0.0598 | 4.346 | 0.000 |
| Central Italy * #70+ | 0.1404 | 0.0703 | 1.997 | 0.046 |
| Southern Italy * #70+ | 0.2922 | 0.0627 | 4.661 | 0.000 |
| Central Italy * Educ. $\geq 8$ | -0.0333 | 0.0289 | -1.150 | 0.250 |
| Southern Italy * Educ $\geq 8$ | 0.0367 | 0.0252 | 1.455 | 0.146 |
| Central Italy * Educ $\geq 13$ | 0.0576 | 0.0307 | 1.873 | 0.061 |
| Southern Italy * Educ $\geq 13$ | 0.0259 | 0.0265 | 0.977 | 0.329 |
| Central Italy * Un. degree | 0.0003 | 0.0479 | 0.007 | 0.995 |
| Southern Italy * Un. degree | -0.0217 | 0.0423 | -0.514 | 0.607 |
| Central Italy * Sex | -0.0114 | 0.0281 | -0.407 | 0.684 |

On the basis of the evidence so far presented, we shall use statistical matching methods to generate a new measure of non-durable expenditure for the SHIW sample. Given the availability of good expenditure data in the SFB, it seems natural to use for this purpose estimates of an (inverse) Engel curve (estimated according to the (3.3) specification). We first create a household-specific estimate of food consumption for all data points in the Bank of Italy SHIW sample generating random drawings from the estimated Gamma distribution on the SFB data set and keeping the first $m$ that fall within the admissible region for household $h$. This imputed value we denote as $food_j^*$ (where index $j$ denotes the $j-th$ imputation).

It is worth stressing that our problem is not a standard prediction problem, because we do not observe over the prediction sample a key explanatory variable, $\ln(food)$, but only its imputed value, $\ln(food)^*$. Even if this is an unbiased predictor, it does not necessarily correlate with the equation error in the same way as $\ln(food)$. For this reason we shall compare our structural form prediction results with the robust, but potentially less efficient, standard reduced form predictions that rely only upon estimated of the first equation in (3.1).

To be more specific, we estimate by OLS on the SFB sample the following two equations:

$$\ln nd = X\pi_1 + Z\pi_2 + u \tag{6.1}$$

$$\ln nd = \gamma_1 \ln(food) + X\gamma_2 + Z\gamma_3 + \omega \tag{6.2}$$

where (6.1) is the reduced form equation used in the standard matching problem (as in Angrist and Kruger (1992), and Arellano and Meghir (1992)), whereas (6.2) is the structural form equation corresponding to the inverse Engel curves discussed in the previous section.

Predictions for the SHIW sample are based upon common information on $X$ and $Z$ in the former method, and the prediction error variance is computed in the standard way. Predictions for the SHIW sample based upon (6.2) are less straightforward: for each household in SHIW we use the common information on $X$ and $Z$ and the $m$ $(= 100)$ different imputations for $\ln(food)$. The prediction error variance must take into account the multiple imputation nature of the exercise, as detailed in the Appendix.

The former method neglects information on food expenditure (when we estimate equation (6.1) we obtain $R^2 = .4156$), is robust to potential misspecification in the imputation procedure and is unaffected by imputation variability. The latter method is potentially more efficient (when we estimate equation (6.2) we obtain $R^2 = .6412$) but it relies on our ability to correctly predict food expendi-

ture in the SHIW sample and its precision is affected by the random nature of the imputation procedure.

We shall show how the imputations based on (6.1) and (6.2) differ. In both cases, we construct the following approximation to the saving rate:

$$s = \ln(y - rent) - \widehat{\ln nd} \tag{6.3}$$

where *rent* is actual rent paid by tenants and imputed rent for home-owners. Our definition treats rent as a pre-committed item of expenditure, that can be taken as fixed in the short run. Also, it implicitly includes in saving total spending in durable goods, and is therefore an upper bound for actual saving.[12]

A saving measure like (6.3) can be constructed on National Accounts data, by taking logarithms of the arithmetic averages of income and expenditure, if we are prepared to include in rent some other items of housing expenditure. In 1995 the saving rate thus defined was 23%. The corresponding statistic in the SHIW data (using the arithmetic averages of reported income, rent and expenditure on non-durable goods) is 47%, an implausibly large number. In fact, the extremely high saving rates implied by the survey have been noted in the literature, and help motivate this paper.

In Table 6.1 we show how the saving rate (6.3) varies across the population according to our choice for $\widehat{\ln nd}$ . In the last column we report the distribution of saving rates based upon observed expenditure. The median is 32%, and the mean is 34%. This compares to the 47% reported above, and suggests that there is less variability in expenditure than in income in the survey.

We can compare these numbers with those obtained when we take imputed measures for $\widehat{\ln nd}$ . If we follow the standard reduced form (RF) procedure, we find a median saving rate of 13.6% and a mean saving rate of 8.6%. The structural form (SF) procedure produces even lower figures (6.6% and 5.6% respectively). As noted in Brandolini and Cannari (1984), income also suffers from underreporting in SHIW, and this makes a straight comparison with national accounts data difficult.

In Table 6.1 we also report standard errors for the mean saving rate, based upon prediction error variances. With the RF method the prediction error variance is the sum of the variance of the disturbance and the variance of the parameter estimates, as usual. With the SF method a third variance component comes

[12] Our measure is not defined for those households whose income net of rent is negative or zero. In our sample this is a relatively rare occurrence (approximately 1% of the whole sample). An alternative that copes well with negative income observations is proposed by Attanasio (1998): $\frac{income-consumption}{consumption}$, where in our case $consumption = \widehat{nd} + rent$. It is a monotonic transformation of the standard measure for all observations with positive income, and it conveys interesting information for the remaining observations.

Table 6.1: Saving rates descriptive statistics

| **Percentiles** | **Estimated (RF)** | **Estimated (SF)** | **Observed** |
|---|---|---|---|
| 5% | -0.8769 | -0.7569 | -0.3646 |
| 25% | -0.1999 | -0.2395 | 0.0847 |
| 50% | 0.1355 | 0.0664 | 0.3202 |
| 75% | 0.4529 | 0.3827 | 0.6226 |
| 95% | 0.9078 | 0.8804 | 1.1281 |
| mean | 0.0862 (0.0063) | 0.0562 (0.0074) | 0.3401 |

into play reflecting the variability induced by the imputation procedure (see the Appendix for further details). The standard errors are of comparable magnitude, but the SF standard error is larger than the RF standard error, indicating that imputation variance is of non-negligible magnitude. The predicted mean saving rates are significantly different from each other, as long as their covariance is null or positive.

In figure 6.1 we show the cumulative density functions for the saving rate corresponding to the observed and the two predicted measures of $\ln(nd)$. The observed cdf lies entirely at the right of the other two, largely because it is based on much lower values for non-durable expenditure. The reduced form and structural form cdf's of the saving rates are close to each other and cross twice, partly reflecting the higher variance of the reduced form saving rate. It is worth noting that the reduced form saving rate is much more skewed, with a relatively fat tail at the left of the support (large negative values).

Much of the literature on savings is interested in the age profile of the saving rate, as the leading economic theory (the life-cycle model of consumption) predicts a hump-shaped age profile for individual households. It is well known that cross-section profiles do not correct for cohort effects, and therefore may provide a misleading picture of the true underlying age profile for each cohort. However, the cross section plot of our different measures may still be interesting if we believe cohort effects to be unaffected in imputation.

In Figure 6.2 we show age plots of observed and imputed saving rates obtained by grouping households in three-year head-age bands. Both predicted measures are consistently below the measure based on observed expenditure, but the age profile is most steeply ascending with the reduced form imputation. Some of the most striking differences that occur at early and late ages may be due to sampling variability (sample sizes are relatively small for ages up to 30 and above 75) but the difference in underlying patterns is likely to reflect the different information used by the two models.

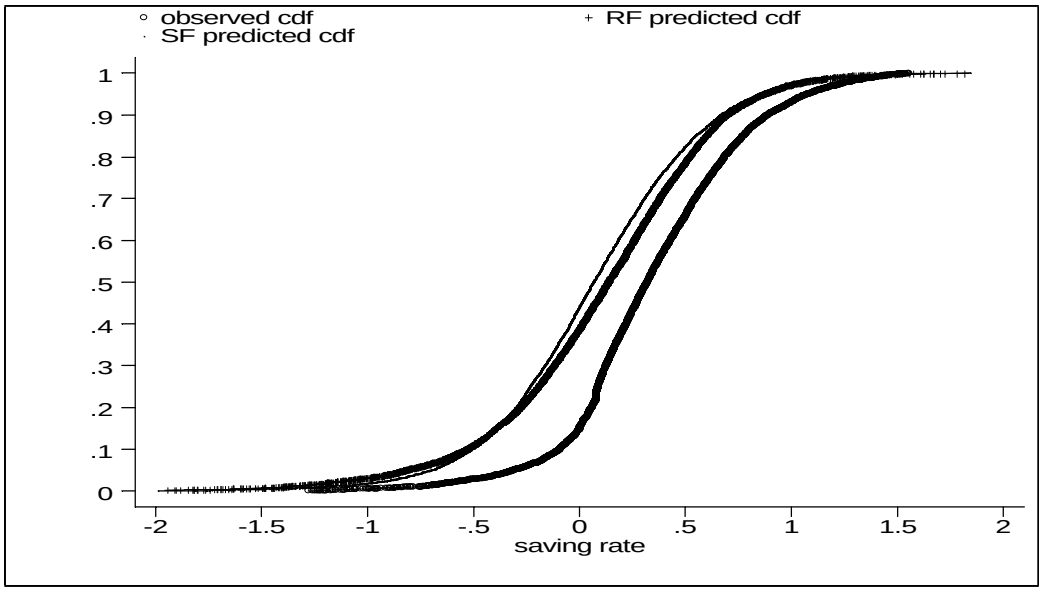


Figure 6.1: Cumulative Density Functions of the Saving Rate

## 7. Conclusions

In this paper we compare food and non-durable expenditure data across two Italian surveys: the widely-used, recall-questions-based Survey of Household Income and Wealth (SHIW) and the newly released diary-based Survey of Family Budgets (SFB). The former contains excellent income and wealth information, but only a few, broad consumption questions; the latter contains detailed records on consumption, but little (if any) income and no wealth information. The two surveys share information on social and demographic household characteristics.

Household-level saving rates based on SHIW information are extremely high for all ages, peaking around or even after retirement age. In this paper we have argued that they are questionable because of the non-standard nature of recall measurement error and that a matching technique should be used to generate predictions for total non-durable expenditure in SHIW, and hence for the saving rate.

In a first step we have analyzed the nature of the recall error process. When we compare marginal densities for food expenditure and total non-durable expenditure, modelling the heaping and rounding process as a function of observed characteristics and the true expenditure level, we find that:

- the SHIW reported food expenditure measure is comparable to the SFB measure once heaping and rounding errors are taken into account

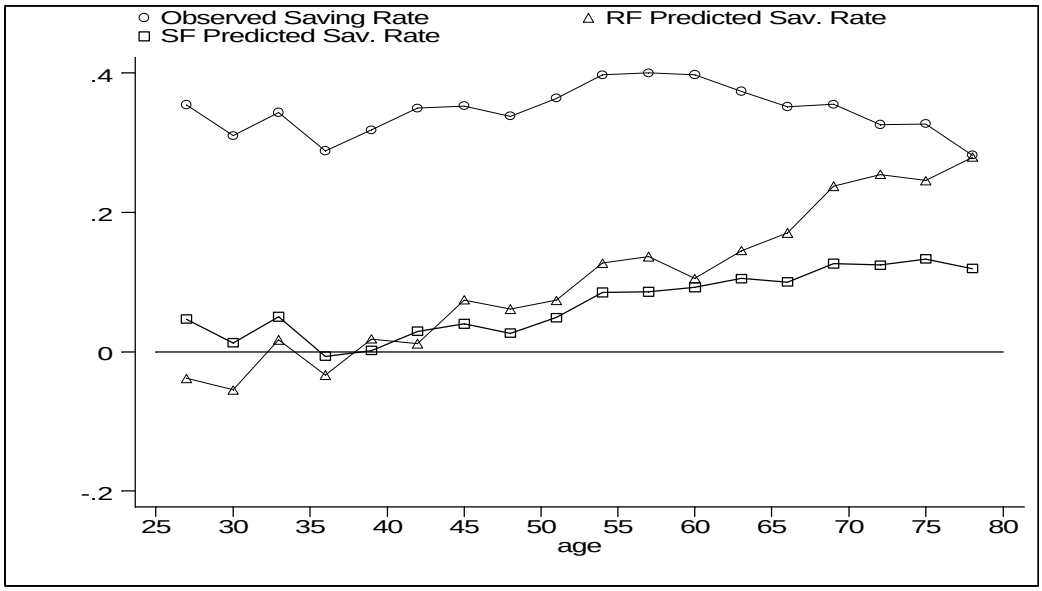


Figure 6.2: Age profiles for the saving rate

- the SHIW reported non-durable expenditure measure is instead more seriously affected by recall error.

On the basis of the above, we have argued that it makes sense to use inverse Engel curve estimates from the SFB to generate an imputation for non-durable expenditure in SHIW. We show that on the SFB sample the non-parametric and semi-parametric double-log Engel curve regressions are very close to straight lines whether we do or do not condition on demographics. We therefore estimate them parametrically by OLS and IV and show to what extent these estimates agree with standard findings on consumer behavior.

We finally have discussed and assessed the relative merits of two prediction techniques: the standard reduced form method that makes no use of food information in the SFB sample and a structural form method that uses food records from both SFB and SHIW samples. This latter method exploits information on reduced form variables and from imputations on SHIW food consumption that are consistent with the estimated heaping and rounding process. Even though more information is used in estimation, the overall precision of the structural form predictions is potentially reduced because of imputation errors. We show that saving rates based on either method are on average much lower than saving rates based on raw data and that their estimated standard errors are of comparable magnitude. The key difference lies in the way they vary with age: the reduced form method generates a markedly upward sloping age profile for the saving rate that is hard

to reconcile with commonly accepted theories on saving behavior; the structural form method produces a much flatter profile. In both cases however there still is evidence of active saving behavior after retirement.

## 8. Appendix

In this section we formally derive the asymptotic standard errors for mean predicted saving rates reported in Table 6.1. Let

$$\widehat{\mathbf{y}}_j = \widetilde{\mathbf{X}}_j \widehat{\boldsymbol{\beta}} \tag{8.1}$$

be the predicted non durable expenditure based on the *j-th* imputed food measure in the SHIW sample, where $j = 1, \ldots, m$. The matrix $\widetilde{\mathbf{X}}_j$ contains the observed SHIW information about household demographic and housing stock characteristics considered in the structural equation estimated in the SFB sample (i.e.: equation 6.2), together with the SHIW j-th imputed measure of food obtained as explained in Section 4.

We predict non-durable expenditure in SHIW for each of the $m$ sets of regressors $\widetilde{\mathbf{X}}_j$ and combine the results to produce final estimate properly adjusted for the multiple imputation context.

Conditional on the generic imputation, standard econometric results can be applied to prove that the forecast in (8.1) is a normal random variable whose variance matrix can be easily estimated by

$$\widehat{\boldsymbol{\Psi}}_j = \widehat{\sigma}^2 \left[ \mathbf{I} + \widetilde{\mathbf{X}}_j \left( \mathbf{X}'\mathbf{X} \right)^{-1} \widetilde{\mathbf{X}}_j' \right].$$

The term $\widehat{\sigma}^2 \left( \mathbf{X}'\mathbf{X} \right)^{-1}$ is the best unbiased estimate for the variance matrix of $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ obtained from the structural equation estimation step in the SFB sample. The estimated variance of the mean predicted expenditure

$$\widehat{\mu}_j = \frac{1}{n} \sum_{i \in SHIW} \widehat{y}_{ij}$$

can be written as

$$\widehat{\tau}_j^2 = \frac{1}{n^2} \mathbf{1}' \widehat{\boldsymbol{\Psi}}_j \mathbf{1}.$$

Pooling the information from the $m$ imputed data sets, the combined estimate and its associated variance are

$$\begin{aligned} \overline{\mu}_m &= \frac{1}{m} \sum_j \widehat{\mu}_j, \\ T_m &= \overline{W}_m + \frac{m+1}{m} B_m, \end{aligned} \tag{8.2}$$

where

$$\begin{aligned} \overline{W}_m &= \frac{1}{m} \sum_j \tau_j^2, \\ B_m &= \frac{1}{m+1} \sum_j \left( \widehat{\mu}_j - \overline{\mu}_m \right)^2 \end{aligned}$$

are the *within* imputation and the *between* imputation sources of variability, respectively, and $(m+1)/m$ is an adjustment for finite $m$. The variance for predicted non durable expenditure (8.2) allows to compute the asymptotic standard errors of our approximation for the saving rate defined in (6.3). Interval estimation and significance tests are based on the statement

$$(\overline{\mu}_m - \mu)\, T_m^{-1/2} \sim t_\rho,$$

where $t_\rho$ is the $t$ reference distribution with

$$\rho = (m-1)\left[1 + \frac{1}{m+1}\frac{\overline{W}_m}{B_m}\right]^2$$

degrees of freedom (see Little and Rubin, 1987, for more details).

## References


[1] Angrist, Joshua D., and Alan B. Krueger (1992) 'The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables With Moments From Two Samples', *Journal of the American Statistical Association,* **87**, 418, 328-336

[2] Arellano, Manuel and Costas Meghir (1992) 'Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets', *Review of Economic Studies*, **59**, 537-557

[3] Attanasio, Orazio (1998) 'Cohort Analysis of Saving Behavior by U.S. Households', *Journal of Human Resources*, 33(3), 575-609

[4] Baldini, Massimo (1998) 'Microsimulazione del comportamento di consumo e integrazione delle indagini campionarie Istat e Banca d'Italia', mimeo, Prometeia

[5] Blundell, Richard, Alan Duncan and Krishna Pendakur (1998) 'Semiparametric Estimation and Consumer Demand', *Journal of Applied Econometrics*, 13, 435-461

[6] Brandolini, Andrea (1998) 'The Personal Distribution of Incomes in Post-War Italy: Source Description, Data Quality and the Time Pattern of Income Inequality', mimeo, Banca d'Italia

[7] Brandolini, Andrea, and Luigi Cannari (1994) 'Methodological Appendix: The Bank of Italy's Survey of Household Income and Wealth', in Ando, Albert, Luigi Guiso and Ignazio Visco (eds.) *Saving and the Accumulation of Wealth. Essays on Italian Households and Government Saving Behavior*, Cambridge: Cambridge University Press

[8] Brugiavini, Agar (1996) 'Un confronto preliminare fra la rilevazione Istat e la rilevazione Banca d'Italia sui consumi delle famiglie', mimeo, University of Venice

[9] Deaton, Angus and John Muellbauer (1980) *Economics and Consumer Behavior,* Cambridge: Cambridge University Press.

[10] Dehejia, R. and Wahba, Sadek (1999), 'Causal effects in non-experimental studies: Reevaluating the evaluation of training programs', *Journal of the American Statistical Association*, **94** (448) : 1053-1062

[11] Fan, J., and I. Gijbels (1996) *Local Polynomial Modelling and Its Application*, Chapman&Hall, London.

[12] Hausman, J. (1978) 'Specification Tests in Econometrics', *Econometrica*, 46(6), 1251-1271.

[13] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998) 'Characterizing Selection Bias Using Experimental Data', *Econometrica*, 66(5), 1017-1098.

[14] Heitjan, Daniel F., and Donald R. Rubin (1990) 'Inference From Coarse Data Via Multiple Imputation With Application to Age Heaping', *Journal of the American Statistical Association*, **85**, 410, 304-314

[15] Heitjan, Daniel F., and Donald R. Rubin (1991) 'Ignorability and Coarse Data', *The Annals of Statistics*, **19**, 4, 2244-2253

[16] Johnson, Norman L., Samuel Kotz and N. Balakrishnan (1994) *Continuous Univariate Distributions* , Wiley, New York

[17] Little, Roderick J.A. and Rubin, Donald B. (1987) *Statistical Analysis with Missing Data*, New York, Wiley

[18] Pesaran, M. Hashem and Richard J. Smith (1994) 'A Generalized $R^2$ Criterion for Regression Models Estimated by Instrumental Variables Method', *Econometrica*, **62, 3**, 705-710

[19] Rosati, Nicoletta (1999) 'Matching statistico tra dati Istat sui consumi e dati Bankitalia sui redditi per il 1995', Economics Department Discussion Paper, **7**, Padua University

[20] Rosenbaum, Paul R., and Donald R. Rubin (1983) 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika*, **70**, 1, 41-55

[21] Torelli, Nicola and Ugo Trivellato (1993) 'Modelling Inaccuracies in Job-Search Duration Data', *Journal of Econometrics*, 59(1/2), 187-211