The effects of income imputation on micro analyses: Evidence from the ECHP*

Cheti Nicoletti
ISER, University of Essex

Franco Peracchi
University of Rome "Tor Vergata"

April 14, 2004

JEL codes: C33, C35, C81

Keywords: Panel data, item nonresponse, imputation, earnings, poverty.

Abstract

Social surveys are usually affected by item and unit nonresponse. Since it is unlikely that a sample of respondents is a random sample, social scientists should take the missing data problem into account in their empirical analyses.

Typically, survey methodologists try to simplify the work of data users by "completing" the data, filling the missing variables through imputation. The aim of this paper is to give data users some guidelines on how to assess the effects of imputation on their micro-level analyses. We focus attention on the potential bias caused by imputation in the analysis of income variables and poverty measures. We consider two methods for evaluating the effects of imputation, using the European Community Household Panel as an illustration.

* Preliminary. Comments welcome.

1 Introduction

Social surveys are usually affected by nonresponse: either failed contact or refusal to fill the questionnaire (unit nonresponse), or refusal to answer specific survey questions (item nonresponse). If the data are missing completely at random (MCAR), that is, the probability of nonresponse does not depend on any observed or unobserved variable, then it is possible to make correct inference about population parameters by considering only the subsample of respondents.

The MCAR assumption is far stronger than necessary, however. In practice, it can often be replaced by the weaker assumption of missing at random (MAR).[1] While this assumption allows the response probability to depend on observed variables, it imposes independence between the response probability and the unobserved variables given the observed ones. The MAR assumption is important because it underlies most imputation procedures employed by survey methodologists to fill the missing data.

The availability of an easy-and-ready-to-use dataset is clearly attractive to most researchers, whose main aim is typically far away from understanding the response behavior of the sample units. Unfortunately, imputation procedures may be inadequate to correct missing data problems, either because they are improperly applied, or because there are too few variables observed for both respondents and nonrespondents that can be used to compute the imputed values.

The aim of this paper is to illustrate how to evaluate the effects of imputation in micro analyses. For concreteness, we consider the effects of imputation on the analysis of income variables and poverty measures in Europe using the European Community Household Panel (ECHP), a longitudinal household survey covering all countries of the European Union.

Much of the literature about imputation focuses on the problem of underestimation of the sampling variance of estimates computed using imputed values (see for example Rubin 1989 and 1996). In this paper we instead focus on the potential bias of estimates of a micro model of interest. We carry out two types of analysis. The first compares various summaries of the distribution of the income variables of interest, in levels and growth rates, between different types of respondents and nonrespondents, using the imputed values given by the ECHP. In particular, we focus attention on conditional models for household income and personal earnings, and show how to check whether relevant variables have been omitted from the imputation procedure. We also show how the presence of imputed values may affect the analysis of earnings dynamics. The second uses instead partial information on income variables to identify upper and lower bounds for the poverty rate. We show how imposing some weak and sensible assumptions helps narrowing the "worst case"

[1] We refer to Rubin (1976) and Little and Rubin (1987) for a formal definition of MAR and MCAR.

bounds originally introduced by Manski (1989). We also suggest an informal test of consistency of imputation. The test simply checks whether the estimated poverty rate lies inside the bounds obtained by imposing weak and sensible assumptions.

The remainder of the paper is organized as follows. Section 2 describes briefly the ECHP, defines the different types of nonresponse affecting income variables, and gives detail on the imputation procedures adopted. Section 3 assumes that MAR holds and shows how to assess consistency of imputation procedures for a specific model of interest. More precisely models for the earnings structure and earnings growth are considered in Sections 3.2 and 3.3 respectively. Section 4 shows how to identify bounds for the poverty rate without imposing MAR. Section 4.1 explains how to use partial reported income to narrow the bounds. Section 4.2 suggests some sensible assumptions helping in narrowing further the bounds. Section 4.3 shows how to check if the poverty rates estimated using the imputed values lie inside the bounds. Finally, Section 5 offers some conclusions.

2 Nonresponse and imputation in the ECHP

This section briefly describes the ECHP,[2] defines the different types of nonresponse for household and personal income, and gives some details on the imputation procedures adopted in the ECHP.

2.1 Brief description of the ECHP

The ECHP is a longitudinal survey of households and individuals, centrally designed and coordinated by the Statistical Office of the European Communities (Eurostat), and conducted annually between 1994 and 2001. Its target population consists of all individuals living in private households within the European Union (EU). In its first (1994) wave, the survey covered about 60,000 households and 130,000 individuals in 12 countries, namely Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain and the UK. Austria, Finland and Sweden began to participate to the ECHP only later, respectively from the second (1995), third (1996) and fourth (1997) wave. In Belgium and the Netherlands, the ECHP was linked from the beginning to already existing national panels. In Germany, Luxembourg and the UK, instead, the first three waves of the ECHP ran parallel to already existing national panels, respectively the German Social Economic Panel (GSOEP), the Luxembourg Social Economic Panel (PSELL) and the British Household Panel Survey (BHPS). Starting from the fourth (1997) wave, it was decided to merge the ECHP into the GSOEP, the PSELL and the BHPS. In this paper, we focus attention on 11 countries which participated to the survey for the first 5 waves, namely Belgium, Denmark,

[2] We refer to Peracchi (2002) for a more detailed description.

France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, and UK.[3]

The ECHP divides the population into sample and nonsample persons. Sample persons are all individuals belonging to the sample drawn for each EU country in the first year of participation, plus children born after the first wave to a sample woman. Nonsample persons are all other individuals. Sample and nonsample persons may or may not be eligible for interview. Sample persons are eligible if they belong to the target population (that is, they live in a private household within the EU) and are aged 16+. In addition, eligibility of nonsample persons also requires them to live in a household containing at least one sample person. Sample persons who are ineligible (homeless, institutionalized, or living outside the EU) are "traced" and interviewed again if they return to the target population. Ineligible nonsample persons are not traced. Sample and nonsample persons whose refusal to respond is considered "final" or did not return a complete questionnaire in two consecutive waves are dropped from the sample. Households not interviewed in two consecutive waves are also dropped.

The ECHP is carried out by National Data Collection Units (NDU), with Eurostat providing centralized support and coordination. The NDUs are responsible for sample selection, adaptation of the questionnaire, fieldwork, basic data processing and editing, and initial weighting of the data. Although Eurostat sets general guidelines in order to ensure comparability of survey results, the NDUs largely rely on their normal rules and routines. All national samples are selected through probability sampling. Procedures are not standardized across countries, however, and each NDU relies on its own methods. In most countries, the sampling frame is either the population register or a master-sample created from the latest population census. The most common sampling procedure is two-stage sampling, with geographical areas (usually the municipalities) as primary sampling units, and households or street addresses as secondary sampling units.

An essential feature of the ECHP is the adoption of a common questionnaire centrally designed by Eurostat. The questionnaire consists of a household register, mainly for record keeping and control of the sample, a household questionnaire submitted to a "reference person" (usually the household head or the spouse/partner of the head), and a personal questionnaire submitted to all eligible household members. The interviewing method recommended by Eurostat is face-to-face personal interviewing, but other interviewing methods have also been used (e.g. telephone or proxy interview). In Greece, Netherlands, Portugal and the UK, interviews are carried out, at least partly, using computer assisted personal interviewing (CAPI). All other countries rely instead mainly on the conventional "paper and pencil" method.

[3] For Germany and the UK we always use the dataset derived form the GSOEP and the BHPS. Luxembourg is excluded because the data for its harmonized national panel are not available in the 2002 release of the data.

2.2 Definition of income components

The income information provided by the ECHP generally consists of annual amounts in the year before the survey, net of taxes and social security contributions, and expressed in national units and current prices.[4] There are exceptions to this general rule for some income sources and some countries. In particular, for Finland and France all income components are collected and reported as gross. Further, the ECHP also provides information on monthly wage and salary earnings (both gross and net) and on total monthly net household income in the month before the survey.[5]

The ECHP distinguishes between six main income sources: wages and salaries, income from self-employment or farming, pensions (old-age related benefits and survivors' benefits), unemployment/redundancy benefits, other social benefits or grants (family-related allowances, sickness/invalidity benefits, education-related allowances, other personal benefits, social assistance, housing allowances), and nonwork private income (capital income, property/rental income, private transfers received). Each income component generally consists of subcomponents, with varying level of detail.[6] Although the questionnaire is very detailed, much of this detail is lost in the process of anonymizing the information and harmonizing the definition of income variables across countries. All income subcomponents are collected at the personal level with the exception of "assigned income" (namely social assistance, housing allowances and property/rental income), which is only collected at the household level and then divided equally among the adult members of a household.

Total personal income is the sum of all personal income components, whether directly collected or "assigned". Personal income components are aggregated at the household level to obtain corresponding household variables. Finally, total household income is obtained by summing over the different types of income and over the individuals belonging to the same household. In what follows, by total net household income (henceforth "household income" for brevity) we mean the sum of personal net incomes of all household members in the year before the survey.

2.3 Income nonresponse

We now turn to nonresponse to income variables for responding households, namely those where at least one eligible member returned the personal questionnaire. We do not consider nonresponding

[4] To allow comparability over time and across countries, all income variables in this paper have been converted to 1995 prices and a common scale by using purchasing power parities.

[5] The latter is available for all countries, including Finland and France.

[6] For example, wage and salary earnings are the sum of regular earnings and lump-sum payments. The latter are the sum of profit-sharing bonuses and other lump-sum payments.

households, for which no data are collected.[7] Within a responding household, we allow for unit nonresponse, that is, we allow for the case when an eligible household member does not return the personal questionnaire.

We also consider item nonresponse, which occurs when an eligible person returns the personal questionnaire but fails to respond to a specific income question. In the case of income aggregates obtained by adding up different components, two types of item nonresponse may arise: full and partial. The former occurs when all income components needed to compute the aggregate are missing (for example, both the regular and the lump sum components of wage and salary earnings are missing). The latter occurs when only some components are missing (for example, regular wage and salary earnings are available but the lump sum component is missing).

In the case of nonresponse to household income, we further distinguish between full and partial income nonresponding households. The former are households where at least one eligible member returns the questionnaire but none answers any income question. The latter consist of (i) households with only item nonresponse (those where all eligible members return the questionnaire but some do not answer all income questions), (ii) households with only unit nonresponse (those where only some eligible members return the questionnaire but they answer all income questions), and (iii) households with both unit and item nonresponse (those where only some eligible members return the questionnaire and not all of them answer all income questions).

2.4 Income imputation

In this section we describe the imputation procedures and the information available in the User Data Base (UDB) of the ECHP to identify unit and item nonresponse.[8]

In the case of item nonresponse to income questions, Eurostat applies an imputation procedure at the individual level to replace the missing personal income components. The procedures adopted have changed through time and some details are provided in Appendix A.

Given the procedures for imputing income at the personal level, the way in which household income is computed depends on the presence of unit nonresponse within a household. For households without unit nonresponse, namely those where all eligible members returned their questionnaire, household income is simply obtained by adding up the reported or imputed values of personal

[7] The ECHP takes into account the presence of nonresponding households by computing weights.

[8] The UDB is an anonymized and user-friendly version of the data. The first release of the UDB, covering waves 1 and 2, was issued by Eurostat in December 1998, three years after completion of fieldwork for wave 2. The second release, covering the first three waves, was issued in December 1999. The third one, covering waves 1–4, was released in June 2001. The fourth one (2002 UDB), covering waves 1–5, was released in February 2002. The fifth wave (2003 UDB), covering waves 1–7, was released in June 2003. In this paper we use only the first 5 waves to assess the imputation changes occurred between the 2002 and 2003 UDBs.

income components.

For households with unit nonresponse, household income is obtained in two steps. In the first step, "imputed household income" Y_h^I is computed as the sum of the reported and imputed incomes components of responding household members, that is,

$$Y_h^I = \sum_i \sum_j D_{hi}[R_{hij}Y_{hij} + (1 - R_{hij})\widehat{Y}_{hij}],$$

where $\sum_i$ denotes summation over all eligible members of household h, $\sum_j$ denotes summation over all subcomponents of income, D_{hi} equals 1 if individual i returns the questionnaire and 0 otherwise, R_{hij} equals 1 if individual i answers the question on the jth subcomponent of personal income and 0 otherwise, and Y_{hij} and $\widehat{Y}_{hij}$ are respectively the observed and imputed jth subcomponent of personal income.

In the second step, "final household income" Y_h^F is obtained by correcting imputed household income Y_h^I for unit nonresponse. The nature of this correction has changed over time. In the 2002 UDB, it consists of inflating imputed household income Y_h^I through a "within-household nonresponse inflation factor" $f_h > 1$.[9] In the 2003 UDB, the correction consists of adding to Y_h^I an "additional income amount", whose computation exploits information on income variables observed in the last wave.[10]

Unfortunately, the UDB provides no flag for income imputation at the individual level. At the household level, two pieces of information are provided. One is the imputation index for item nonresponse, defined as

$$W_h = 1 - \frac{\sum_i \sum_j D_{hi}R_{hij}Y_{hij}}{Y_h^I} = 1 - \frac{Y_h^R}{Y_h^I},$$

where $Y_h^R = \sum_i \sum_j D_{hij}R_{hij}Y_{hij}$ is reported household income. For the 2002 UDB, the other piece of information is the within-household nonresponse inflation factor f_h, whereas for the 2003 UDB it is the "additional household income". Both indices are only available at the household level and do not give enough information to distinguish between reported and imputed income at the personal level. This distinction is however possible for households with a single recipient of the income category of interest. Further, for households with more than one income recipient, the two indices provide information useful to distinguish between the four types of income nonresponse at the household level introduced in Section 2.3 (full income nonresponding households, households

[9] The within-household nonresponse inflation factor is constant within households and for all the subcomponents of income. For example, the household income from self-employment and wages and salaries after item imputation are multiplied by a same factor to take account of unit nonresponse, it does not matter whether the unit nonrespondents were working or not working as self-employed or employees last year.

[10] See Appendix B for some details on the computation of f_h and of the additional household income.

with only item nonresponse, households with only unit nonresponse, and households with both unit and item nonresponse).

2.5 Imputation of household income

Table 1 shows the importance of item and unit nonresponse using three different concepts of household income, namely reported income Y_h^R (the sum of the personal incomes reported by each household member), imputed income Y_h^I (the sum of reported and imputed personal incomes), and final income Y_h^F (imputed household income multiplied by the within-household inflation factor of adjusted by adding the additional household income). The table also shows the item imputation ratio defined as the ratio Y_h^R/Y_h^I between the reported and the imputed household income (and therefore equal to $1 - W_h$), the unit imputation ratio defined as the ratio Y_h^I/Y_h^F imputation ratio defined as the ratio Y_h^R/Y_h^F between the reported and the final household income (and therefore equal to the product of the two previous ratios). All three ratios vary between zero and one, with zero corresponding to cases when the entire household income comes from imputation (item, unit or total) and one corresponding to cases when no imputation takes place.

We are interested in checking whether there are systematic differences in the distribution of household income across the different types of responding households defined in Section 2.3 (full income nonresponding, only item nonresponding, only unit nonresponding, and both unit and item nonresponding households). Table 1 shows the average values of household income and of the imputation indices for the four types of responding households. The columns correspond to the different types of responding household, while the rows correspond to the variables (imputation indices and household incomes) for which the average is computed. The table is divided in two parts: the top part shows the results using the 2002 UDB, while the bottom part shows the results for the 2003 UDB.

Item imputation is more important than unit imputation. About 22% of the households have problems of item nonresponse for their members, but only 3%–4% of them have problems of unit nonresponse. The last column of Table 1 shows that the average value of the item imputation ratio (.934 and .936 respectively for the 2002 and the 2003 UDB) is smaller than that of the unit imputation ratio (.985 and .988 respectively for the 2002 and the 2003 UDB). This implies that, on average, reported household income is inflated by 1.2–1.5% due to unit imputation, and by 6.4–6.6% due to item imputation.

In the case of full item nonresponse, the imputation procedure in the ECHP delivers a lower average household income for nonrespondents than for respondents. To check whether the differences in final household income across types of responding household depend on the different

characteristics of these households, we estimate a median regression model whose covariates consist of the age of the reference person, the size of the household, the number of children aged less than 16, the number of workers, a dummy for a reference person without a spouse, and dummies for the schooling attainments of the reference person. The model intercept represents median household income in 1997 of an individual with age equal to the average, married, with secondary education completed, and living in a household whose size, number of workers and number of children are all equal to the average. The intercept is very high for households with both unit and item nonresponse, but quite low for households with full item nonresponse (Table 1). In conclusion, the potential bias of the imputation procedure for household income does not seem to disappear after controlling for the characteristics of a household.

To understand whether relevant information has been excluded from the imputation procedures, we look at the joint significance of the variables in the regression for log household income, estimated separately for the different types of responding households. We report the pseudo R^2 as a measure of the joint significance of the variables in Table 1. In the 2002 UDB, a large fall in the pseudo R^2 is observed for households with unit nonresponse, full income nonresponse, and both unit and item nonresponse relative to those with only item nonresponse. In the 2003 UDB, the fall in the pseudo R^2 is observed instead only for full income nonresponding households.

Looking at the regression for households with only unit nonresponse, we find a considerable increase of the adjusted R^2 in the 2003 UDB relative to the 2002 UDB. Moreover, while in the 2003 UDB the estimated regression coefficients are similar to those obtained for the respondents, in the 2002 UDB the differences are substantial.[11] This evidence suggests an improvement in the unit imputation procedures. In particular, while there is no relationship between imputed household income and household size and educational attainments when using the 2002 UDB, the relationship becomes strong and similar to that estimated for responding households when using the 2003 UDB. An improvement can also be observed when comparing the regression for households with both unit and item nonrespondents, with the R^2 increasing from .21 in the 2002 UDB to .31 in the 2003 UDB.

For full income nonresponding households, the relationship between household income and the explanatory variables is not very strong. Moreover, the value of the coefficients is quite different from the equation for responding households. This reflects a possible problem in the imputation procedure, which seems to underestimate household income for full income nonresponding households.

Income underestimation for fully nonresponding households is as serious in the 2003 UDB as it was in the 2002 UDB. A similar, but slighter, income underestimation problem is present for the

[11] We obtain the same results when considering mean regressions.

households with only unit nonresponse in the 2002 ECHP, but the problem disappears in the 2003 UDB. It seems to us that the new method of correcting for unit nonresponce within responding households works better. In particular, it makes sense to allow for unit nonresponse by computing an "additional income amount" which depends on current and lagged personal and household characteristics, rather than inflating household income by a factor that is completely unrelated to the characteristics of the unit nonrespondents.[12]

If we break down Table 1 by country and by wave, we again find that the average value of final household income for full item nonresponding households is lower than for fully responding households. There are only two exceptions in the 2003 UDB, namely Ireland and the UK, which are incidentally the two countries which provided their own procedures for the 2003 UDB. In the 2002 UDB, the exceptions are instead Belgium and Ireland, where however the number of cases with full item nonresponse are very low.

2.6 Imputation of personal earnings

This section examines imputation of personal earnings, separately for wages and salaries and self-employment income. All quantities are annual net amounts in the year before the survey (except for France where earnings are gross), and are all evaluated at constant 1995 prices and converted to the same scale using purchasing power parities.

To examine imputation at the individual level, we focus on households with a unique earner of the specific income considered. This allows us to use the imputation index, computed at the household level, as an individual imputation index taking value one in case of full item nonresponse, value zero in the opposite case of full response, and values between zero and one in cases of partial item nonresponse. We do not consider households where there are both item and unit nonresponse, and we focus attention only on nonresponse to earnings. By reported and final earnings we mean, respectively, personal earnings before and after the item imputation.

Table 2 shows the incidence of imputation on earnings (wages and salaries and self-employment income) in terms of the percentage of imputed units and the average imputation index. It also shows the mean of final earnings and the mean difference between final and reported earnings. By column, the table reports the above variables computed separately for responding, partially imputed, fully imputed units and all units.

There is a strong association between the type of income and the nature and incidence of item nonresponse. The percentage of nonrespondents is much higher for self-employment income than for wages and salaries. Wages and salaries are mainly affected by partial item nonresponse, while

[12] See Appendix B for more detail.

self-employment variable is affected mainly by full item nonresponse. There is no case of partial nonresponse to self-employment income in the 2003 UDB. Going from the 2002 to the 2003 UDB, the number of respondents increases while nonresponse decrease. Moreover, the number of people who only earn wage and salary income increases, mainly due to changes occurred in Germany and the UK.

3 Assessing imputation procedures assuming MAR

In this section we assess the impact of imputation on estimates of population aspects of interest under the maintained assumption that the data are missing at random (MAR). Specifically, Section 3.2 looks at the impact of imputation on estimates of the cross-sectional stricture of earnings, while Section 3.3 looks at the impact on estimates of earnings growth.

3.1 The MAR assumption

The conditional mean $\mathrm{E}(Y \mid X)$ and the conditional percentiles of income given a set X of observed variables may be consistently estimated using imputed values if the following assumptions hold:

1. The data are MAR, that is, the response probability does not depend on income conditional on a set (X_c, X_y, X_r) of observed variables, where the variables in X_c affect both the distribution of income and the response probability, the variables in X_y affect only the distribution of income, and the variables in X_r affect only the response probability.

2. The imputation procedure properly uses the variables in $X = (X_c, X_y)$ as auxiliary variables to compute the imputed values Y^I, that is, $\mathrm{E}(Y^I \mid X) = \mathrm{E}(Y \mid X)$.

3. The distribution of income depends on X only through the location parameter.

If condition (1) holds, then we can check whether variables in X have been incorrectly omitted by the imputation procedure. It is enough to replace the missing income variables with the imputed ones and then estimate separate regression models for the respondents and the nonrespondents. Under the MAR condition, all variables in X which help explain the variability of income for the respondents should also help explain the variability of income for the nonrespondents. Testing whether there are variables in X that are relevant for the respondents but irrelevant for the nonrespondents helps identify variables that have been omitted or improperly used in the imputation procedure. In conclusion, as long as variables not used in the imputation are irrelevant either for income or for the response probability, the imputation remains consistent. By contrast, when the

imputation procedure does not properly use all the variables in X, estimates computed using the imputed values are no longer consistent.

In both the 2002 and the 2003 UDB, the auxiliary variables used to impute wages and self-employment income are: the region of residence and the number of employed people in the household, the age, gender, schooling level, occupation in the current job, status in employment, job status, and total number of hours worked per week of the person, and the main activity and size of the local unit where the persons is working. For self-employment income, the marital status of the person and the "equivalized" household size (using the modified OECD scale) are also used.

No use is made of variables linked to the data collection process, such as the mode of interview, the indicator for the use of the same interviewer across waves, and the number of visits to the household. This does not cause any problem if these variables are relevant for the response probability model but irrelevant for the model explaining the variability of income variables. This sort of assumption is generally a sensible one, and we will use it in Section 4.2 below.

3.2 Earnings structure

In this section we analyze some features of the distribution of wages and salaries and self-employment earnings. As in Section 2.6, we focus attention on people who are the unique earners, within their household, of the specific type of earnings considered.

Table 3 shows the the mean, the median, the 10th and the 90th percentile of log earnings for three different types of individuals: full respondents, partial item nonrespondents, and full item nonrespondents. We use reported income for the respondents and imputed values for the nonrespondents. Table 4 presents the estimated intercepts and their standard errors for the median, the 10th and the 90th percentile regressions of the logarithm of wages and salaries (or self-employment incomes) on work experience, its square, and indicators for people without a spouse, schooling, sex and wave.

Comparing the averages, the medians, the 10th and 90th percentiles across types of nonresponse in the 2002 UDB (Table 3), it seems that partial item nonrespondents have the highest earnings followed by respondents and full item nonrespondents. After controlling for a set of explanatory variables, however, the differences between partial item nonrespondents, respondents and fully nonrespondents narrow considerably (Table 4). The only exception are full nonrespondents to wage and salary income. Looking at the same marginal and conditional summary statistics in the 2003 UDB, we again find very low average incomes for the full nonrespondents.

Thus, it seems that the imputation procedure adopted to solve the full nonresponse problem produces seriously underestimated values for wages and salaries. However, because the percentage

of full nonrespondents is quite low (.6% in the 2003 UDB and .9% in the 2002 UDB, see Table 2), the bias in the average wage and salary computed using all individuals is likely to be small. On the other hand, although full item nonresponse for self-employment earnings is high (more than 33%), conditional and unconditional mean and percentiles of self-employment income do not differ significantly for respondents and full item nonrespondents (Table 3).[13]

In conclusion, wages and salaries of full item nonrespondents appear to be underestimated. However, the number of cases involved is relatively small, and so statistics computed for the full sample and the subset of respondents do not differ much. For self-employment income, instead, full item nonresponse is very frequent, but we find no evidence of bias.

3.3 Earnings growth

This section addresses the problem of the impact of imputation on estimates of earnings growth. Although the imputation procedures appear to work reasonably well when attention focuses on the structure of earnings in a single wave, their appropriateness may be an issue when focusing on earnings dynamics. Clearly, if a person is item nonrespondent in wave $t-1$ but responds in wave t, then part of her earnings growth may be a consequence of the imputation procedure. In this section we show that this may cause an increase of extreme values, that is, values in the tails of the distribution of earnings growth. We also show that the presence of extreme values affects inference about the mean growth rate of earnings, while inference about the median is more robust.

If we consider the earnings of an individual in two consecutive waves, nine response patterns can be distinguished: (1) response in both waves, (2) response in the first wave followed by partial item nonresponse, (3) response in the first wave followed by full item nonresponse, (4) partial item nonresponse followed by response, (5) partial item nonresponse in both waves, (6) partial item nonresponse followed by full item nonresponse, (7) full nonresponse followed by response, (8) full item nonresponse followed by partial item nonresponse, and (9) full item nonresponse in both waves.

For people who are partial nonrespondents in any of the two periods, the average growth rate of wages and salaries does not differ much with respect to people who are respondents in both periods. On the other hand, those who are full item nonrespondents in wave $t-1$ have larger average earnings growth, whereas those who are full item nonrespondents in wave t have a negative growth rate. Taken together, these two findings may be another signal of underestimation of wages and salaries for full item nonrespondents.

If we look instead at the median growth rate of earnings, estimates change much less across

[13] Repeating the same analysis separately by country gives similar results.

response categories. The interdecile range for item nonrespondents in at least one wave is generally larger than for people with complete response in both waves. In other words, extreme values for the growth rate occur more frequently for full item nonrespondents, thus explaining the instability of the average relative to the median.

Turning to self-employment income, mean growth again varies more than median growth across response categories. Contrary to the case of wages and salaries, however, both the mean and the interdecile range are now higher for full item nonrespondents in wave t.

We have also estimated various quantile regression models for earnings growth by grouping the different types of nonresponse into six categories: respondents, full item nonrespondents and partial item nonrespondents, respectively in wave $t-1$ and wave t. The model specifies the conditional quantile given a vector X of observable individual characteristics as $m(X) = \alpha + \beta^\top X$, where α and β are parameters to be estimated and the vector X includes age, age square, sex, job duration, an indicator for people without a spouse, and indicators for education completed (2 dummies, one for college and one for primary education). The intercept α for the 10th (50th and 90th) percentile regression corresponds to the 10th (50th and 90th) percentile of earnings growth in 1996 for a worker with age equal to the average, married, with an average job duration and with secondary education completed. Table 6 shows the parameter estimates, the standard error (S.E.) of the intercept and the number of observations (no. obs.) for each category of response and the three percentiles. The results confirm the fact that the median changes less than the 10th and 90th percentiles even after conditioning on a set of explanatory variables.

The interdecile ranges are generally shorter in the 2003 UDB than in the 2002 UDB. This may indicate that the problem of extreme values is less serious in the latest ECHP release.

We also checked whether the significance of the explanatory variables changes across responding categories. Although the growth rate of wages and salaries is hard to predict in general, it seems that work experience and educational attainments are significant, except when considering full item nonrespondents in wave t. Predicting the growth rate of self-employment income is even more difficult, and the significance of the predictors is quite low for both the respondents and the nonrespondents.

In conclusion, imputation appears to affect the estimates of the earnings dynamics. In particular, imputation appears to alter the tails of the distribution of earnings growth, increasing the the relevance of extreme cases. An analysis of earning dynamics should take this into account by using statistics that are robust to the presence of extreme values, such as the median instead of the mean, or median regression instead of mean regression.

4 Assessing imputation procedures relaxing MAR

Our aim in this section is to assess the impact of the income imputation procedures adopted in the ECHP on estimates of the poverty rate, a problem of direct importance for some ECHP users.[14] We begin by computing bounds for the poverty rate without imposing any assumption on the missing data process. We then try to narrow the bounds by using the information on reported income available for partial item nonresponding households. Finally, we impose some weak assumptions to further reduce the width of the bounds, and look for cases where the poverty rates computed using the imputed values do not respect the resulting bounds.

4.1 Partial identification of poverty rates

Poverty is conventionally defined as having an income below a "poverty line", usually a fraction of median income. A common income concept is net equivalized household income, obtained by dividing total household income by some equivalence scale. In this paper, we use net household income expressed in thousands Euro at constant 1995 prices, rescaled by the modified OECD equivalence scale. We compute median income and the probability of being poor (the "poverty rate") separately by country using all members (both children and adults) of responding households in the ECHP.

Following Manski (1989), let D be a dummy variable taking value 1 if an individual belongs to a responding household whose total income Y is fully reported and 0 otherwise, and let c be the poverty line. By the law of total probability, the poverty rate $\Pr\{Y \leq c\}$ satisfies

$$\Pr\{Y \leq c\} = \Pr\{Y \leq c \,|\, D = 1\} \Pr\{D = 1\} + \Pr\{Y \leq c \,|\, D = 0\} \Pr\{D = 0\}.$$

Because only three of the four elements on the right hand side of the above expression can be identified (consistently estimated), the poverty rate cannot be identified from data subject to item nonresponse. Notice however that the unknown element $\Pr\{Y \leq c \,|\, D = 0\}$ takes values between 0 and 1. We can therefore compute an upper bound

$$UB = \Pr\{Y \leq c \,|\, D = 1\} \Pr\{D = 1\} + \Pr\{D = 0\}$$

and a lower bound

$$LB = \Pr\{Y \leq c \,|\, D = 1\} \Pr\{D = 1\}$$

for the poverty rate simply by replacing the unknown element with its maximum and minimum values. These bounds are known as "worst case bounds". The range of plausible value for the poverty rate is equal to $UB - LB = \Pr\{D = 0\}$.

[14] For example, the European Commission uses the ECHP to construct indicators of social exclusion.

4.2 Narrowing the bounds

An important question is how to narrow the worst case bounds, that is, how to sharpen our inference by reducing the range of plausible values for the poverty rate.

We begin by noticing that most nonrespondents provide partial information on their income. For example, we typically know reported household income Y^R, which represents a lower bound for household income. We can then decompose the unknown poverty rate as follows

$$\Pr\{Y \leq c \,|\, D=0\} = \Pr\{Y \leq c \,|\, Y^R \leq c, D=0\} \Pr\{Y^R \leq c \,|\, D=0\} + \\ + \Pr\{Y \leq c \,|\, Y^R > c, D=0\} \Pr\{Y^R > c \,|\, D=0\}.$$

Since actual income Y is always greater or equal to reported income Y^R, it follows that $\Pr\{Y \leq c \,|\, Y^R > c, D=0\} = 0$. Because the reported household income of nonresponding individuals is known, $\Pr\{Y^R \leq c \,|\, D=0\}$ can easily be estimated. The probability $\Pr\{Y \leq c \,|\, Y^R \leq c, D=0\}$ is instead unknown, but must necessary lie between 0 and 1. This allows us to compute a new upper bound

$$UB_R = \Pr\{Y \leq c \,|\, D=1\} \Pr\{D=1\} + \Pr\{Y^R \leq c \,|\, D=0\} \Pr\{D=0\}.$$

The information on reported income does not affect instead the lower bound, which remains unchanged with respect to the worst case bound LB. Thus, the use of partially reported income allows us to narrow the width of the bound from $\Pr\{D=0\}$ to $\Pr\{Y^R \leq c \,|\, D=0\} \Pr\{D=0\}$.

Table 7 shows the estimated worst case bounds with their bootstrap confidence intervals (CI lower for the lower bound and CI upper for the upper bound), the estimated upper bound UB_R using reported income with the corresponding boostrap confidence interval (CI_R upper), and the sample size. The top, central and bottom parts of the tables correspond to three different poverty lines, namely 40%, 50% and 60% of median income. Bootstrap confidence intervals are obtained by drawing with replacement 1000 samples from the original data, and then taking the 5th and 95th percentiles of the bootstrap distribution. The results, presented in Table 7 and subsequent tables, are based on the 2003 UDB.[15]

Because the uncertainty due to sampling variability is small relative to the uncertainty due to item nonresponse, the main issue is finding weak assumptions to further narrow the bounds. Using reported income helps, as UB_R is always much smaller than the worst case bound UB.

4.3 Assessing consistency of imputation

As emphasized by Horowitz and Manski (1998), "estimates using imputations take the observed data as given and specify logically possible values for the missing data. Hence imputation always

[15] The results obtained for the 2002 UDB are very similar and are available from the Authors upon request.

yields a logically possible value of the conditional expectation of interest". In our case, imputed household income Y^I clearly satisfies $0 < \Pr\{Y^I \leq c \,|\, D = 0\} < 1$. Because the imputed poverty rate $\Pr\{Y^I \leq c\}$, obtained from the completed data, necessarily satisfies

$$\Pr\{Y^I \leq c\} = \Pr\{Y \leq c \,|\, D = 1\} \Pr\{D = 1\} + \Pr\{Y^I \leq c \,|\, D = 0\} \Pr\{D = 0\},$$

it must always lie between the worst case bounds. The bounds computed using reported income are narrower but, again, they always contain the imputed poverty rate. This is because imputed income is always greater or equal to reported income, which implies that

$$0 < \Pr\{Y^I \leq c \,|\, D = 0\} < \Pr\{Y^R \leq c \,|\, D = 0\}.$$

The imputed poverty rate may instead lie outside the bounds computed by imposing additional assumptions. In what follows, we consider four types of weak assumptions that help narrow the worst case bounds. We then check whether there are inconsistencies in the imputed poverty rates, i.e. cases in which they lie outside the computed bounds.

Our first assumption is that, conditional on a suitable set of variables X, the poverty rate is independent of the indicator for the use of the same interviewer across waves. In other words, we use the above dummy as an instrumental variable (IV). Our second assumption is that, conditional on a set of variables X, the poverty rate is lower for people living in households where the reference person has a higher education level, that is, the indicator for the level of education (1 for primary education and 0 for higher education) is a monotone instrumental variable (MIV) in the sense of Manski (1995) and Manski and Pepper (2000). Under assortative mating between the reference persons and the cohabiting people, the education level of the former is a good proxy for the education level of the other adults living in the household. Our third and fourth assumptions state that, conditional on a set of variables X (including the above IV and MIV), households with a higher number of working members are richer, while larger households are poorer.

If Z is an IV, then $\Pr\{Y \leq c \,|\, X, Z) = \Pr\{Y \leq c \,|\, X)$. In this case, the bounds for $\Pr\{Y \leq c \,|\, X)$ are given by the intersection of the bounds computed for different values of Z, i.e.

$$LB_{IV} = \sup_z \Pr\{Y \leq c \,|\, X, Z = z, D = 1\} \Pr\{D = 1 \,|\, X, Z = z\},$$
$$UB_{IV} = \inf_z \Pr\{Y \leq c \,|\, X, Z = z, D = 1\} \Pr\{D = 1 \,|\, X, Z = z\} + \Pr\{D = 0 \,|\, X, Z = z\}.$$

If Z is instead a MIV, then $\Pr\{Y \leq c \,|\, X, Z = z_1\} > \Pr\{Y \leq c \,|\, X, Z = z_2\}$ whenever $z_1 < z_2$. In this case, the bounds for $\Pr\{Y \leq c \,|\, X, Z)$ are

$$LB_{MIV} = \sup_{z_1 > z} \Pr\{Y \leq c \,|\, X, Z = z_1, D = 1\} \Pr\{D = 1 \,|\, X, Z = z_1\},$$
$$UB_{MIV} = \inf_{z_2 < z} \Pr\{Y \leq c \,|\, X, Z = z_2, D = 1\} \Pr\{D = 1 \,|\, X, Z = z_2) + \Pr\{D = 0 \,|\, X, Z = z_2\}.$$

The IV and MIV bounds are all conditional on a set X of covariates. In order to compute bounds for the poverty rate by country, we integrate out the conditioning variables by using their marginal distribution.

Table 8 presents the percentage of cases where the imputed poverty rate falls outside the bounds obtained by exploiting the information contained in the IV and MIV assumptions. The IV is a dummy variable indicating the use of the same interviewer across waves, MIV1 is the number of workers in the household (0, 1, 2+), MIV2 is a dummy variable indicating the education level of the reference person (1 for third level education and 0 otherwise), and MIV3 is the household size (1–2, 3, 4+)). For each IV and MIV, we consider the worst case bounds and those computed using the information on the partially reported income. The table is again split in three parts corresponding to our three different poverty lines.

Since we do not observe the bounds but only their estimates, we also have to take sampling variability into account. In checking whether the predicted poverty rate lies outside the bounds, we replace the estimated lower bounds with the lower value of their 95% confidence interval and the estimated upper bounds with the upper value of their 95% confidence interval. We consider acceptable a percentage of inconsistencies (cases in which the predicted poverty rate lies outside the bounds) lower or equal to 10%.

When the dummy indicating the presence of the same interviewer across waves is used as an IV, the percentage of cases in which the imputed poverty rate lies outside the bounds is quite high in France, Germany, Ireland, the Netherlands and the UK. For Belgium, Denmark, Greece, Italy and Portugal the percentage is instead always lower than 15%.[16]

Inconsistencies are less frequent for the MIV bounds. When using the number of workers in the household as a MIV, we find that the percentage of inconsistencies is above 10% only in Spain and the Netherlands. When the education level of the reference person is used as a MIV, only the Netherlands present a percentage of inconsistencies higher than 10%, and only when the poverty line is defined as 40% of the median household income. Finally, when the household size is used as a MIV, only France and Portugal present percentages of inconsistencies higher than 10%. Belgium, Greece, Italy and Spain are the countries with fewer inconsistencies. It is instead surprising the large number of inconsistencies observed for France, Germany, Ireland, the Netherlands and the UK when using as an IV the dummy variable indicating continuity of the interviewer. This can be due to the fact that continuity of the interviewer is less likely in areas, such as big cities, where the household income can be on average different from other areas. To check this, we computed the

[16] There are no inconsistencies for the IV bounds in Spain because, for that country, we do not know whether a household has been contacted by the same interviewer across waves.

correlation between the area average of household income and the percentage of cases in which the same interviewer has been used.[17] The correlations are not significantly different from 0 except in Germany and Greece.

Table 9 shows the differences between the upper and the lower bound under the IV and the the MIV assumptions. IV indicates again the dummy for the use of the same interviewer, while MIV1, MIV2 and MIV3 indicate our three alternative MIV: the number of workers, the education level and the household size. Again we present for each IV and MIV the results for the standard bounds and the reported bounds, i.e. the bounds computed using partially reported income. In Table 10 we report instead the estimated poverty probabilities (using the imputed values when income is missing and the observed values otherwise) for the full sample, for people belonging to fully responding households and for people not belonging to fully responding households. We also report the percentage of people with a partial reported income lower than the poverty line, the width $UB - LB$ of the worst case bound, and the width $UB_R - LB$ of the bound computed using partially reported income. Notice that the former width is equal to the probability of nonresponse $\Pr\{D = 0\}$ (or, more precisely, the probability for people belonging to responding households to have full or partial item nonresponse to household income), while the latter is equal to $\Pr\{Y^R \leq c, D = 0\} = \sum_x \Pr\{Y^R \leq c \,|\, D = 0, X = x\} \Pr\{D = 0 \,|\, X = x\} \Pr\{X = x\}$. Again, the top, central and bottom parts of the tables correspond to the three different poverty lines considered in this paper.

The widest bounds are obviously the worst case ones, followed by the MIV and the IV bounds. Using the partially reported income to narrow the bounds helps a lot. The width of the reported bounds increases by choosing a higher poverty line. The countries with a lower poverty rate and a lower nonresponse probability have narrower reported bounds. Because the width of the worst case bounds is equal to the probability of nonresponse, it does not change using different definitions of poverty. When using an IV or a MIV, the width of the bounds does not change significantly by applying different definitions of poverty. The IV and MIV bounds making use of partially reported income are much narrower than the worst case ones.

5 Conclusions

This paper analyzes a number of issues surrounding income nonresponse and income imputation, using the ECHP as an illustration.

Comparing final household income for different types of responding households using the 2002

[17] Unfortunately, area identifiers are not very detailed in the UDB.

and the 2003 UDB, we find that relevant improvements have been made to the imputation procedure in order to take into account unit nonresponse within responding households. While the average final household income seems to be underestimated for households with unit nonresponse in the 2002 UDB, this is no longer true in the 2003 UDB. A similar underestimation problem may be present for households with full income item nonresponse. This problem persists in both the 2002 and the 2003 UDB, but is attenuated by the fact that very few households have problems of full item nonresponse.

Except for the imputation of unit nonresponses, there have been no other changes in the imputation procedures between the 2002 and the 2003 UDB, and the results concerning item nonresponse are similar between the two releases of the data. If we consider the structure of earnings in a single wave, the imputation procedure for item nonresponse seems to work well, with the possible exception of a few cases of full item nonresponse on wages and salaries. For self-employment income, instead, full item nonresponse is very common but does not appear to lead to significant biases. If consider estimates of earnings dynamics, however, we find that imputation may alter the tails of the distribution of earnings growth. This effect may be reduced by considering statistics that are robust to outliers.

The information on partially reported household income plays a very important role in narrowing down the bounds for the poverty rate. The length of the bounds becomes quite small when the poverty rate is low, and it becomes even smaller after conditioning on a set of variables. Obviously, choosing a low poverty line helps reduce the identification problem. Furthermore, when using some weak IV and MIV assumptions, we notice that the poverty rates computed using the imputed values sometimes lie outside the bounds. This may be due to the use of improper instruments, but may also be due to the ECHP imputation procedures.

In conclusion, we suggest using the reported bounds for the poverty rates in addition to the point estimates computed using the imputed values, at least when the analysis of poverty is conditional on a set of variables, which allows to narrow significantly the bounds.

References

Eurostat (2001, 2002), Imputation of income in the ECHP, PAN 164.

Eurostat (2002), Construction of weights in the ECHP, PAN 165.

Heckamn J.J., Lochner L. and Todd P.E. (2001), Fifty years of Mincer earnings regressions, mimeo.

Horowitz J.L. and Manski C.F. (1998), Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputation, *Journal of Econometrics*, 84: 37–58.

Little J.A. and Rubin D.B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Manski C.F. (1989), Anatomy of the selection bias, *Journal of Human Resources*, 24: 343-360.

Manski C.F. (1995), *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Manski C.F. (2003), *Partial Identification of Probability Distributions*, Springer-Verlag, New York.

Manski C.F. and Pepper J.V. (2000), Monotone instrumental variables: With an application to return to schooling, *Econometrica*, 68: 997–1010.

Nicoletti C. and Peracchi F. (2002), A cross-country comparison of survey participation in the ECHP, ISER Working Paper No. 2002–32, University of Essex, Colchester.

Peracchi F. (2002), The European Community Household Panel: A review, *Empirical Economics*, 27: 63–90.

Raghunathan T.E., Solenberger P.W. and Hoewyk J.V. (1999), *IVEware: Imputation and Variance Estimation Software. Installation Instructions and User Guide. Survey Methodology Program*, Survey Research Center, Institute for Social Research, University of Michigan.

Rubin D.B. (1976), Inference and missing data, *Biometrika*, 63: 581-592.

Rubin D.B. (1989), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Rubin D.B. (1996), Multiple imputation after 18+ years, *Journal of American Statistical Association*, 91: 473–520.

Table 1: Average values of imputation ratios and of household income by type of nonresponse.

	Complete response	Type of nonresponse				Total
		Only item	Full item	Only unit	Item/Unit	
2002 UDB						
Item imputation ratio	1	0.791	0.000	1.000	0.636	0.934
Unit imputation ratio	1	1.000	1.000	697	0.662	0.985
Total imputation ratio	1	0.791	0.000	0.697	0.457	0.926
Reported income	9.175	8.495	0	6.121	4.684	8.720
Item imputed income	9.175	10.763	6.266	6.120	7.056	9.293
Final income	9.176	10.763	6.266	8.985	11.225	9.476
Median regression						
Intercept	9.297	10.315	5.182	8.585	10.552	9.475
Standard error	0.012	0.034	0.081	0.572	0.053	0.012
Adjusted R^2	.29	.26	.08	.10	.21	
No. obs.	562,685	118,452	11,652	678	31,229	724,696
Percentage	77.6	16.3	1.6	0.1	4.3	100
2003 UDB						
Item imputation index	1	0.798	0	1	0.605	0.936
Unit imputation index	1	1	1	0.677	0.624	0.988
Total imputation index	1	0.798	0	0.677	0.434	0.931
Reported income	9.137	8.348	0	6.287	4.255	8.700
Item imputed income	9.137	10.344	6.339	6.287	6.195	9.208
Final income	9.137	10.345	6.339	9.943	10.095	9.329
Median regression						
Intercept	9.336	10.282	5.364	10.043	10.078	9.483
Standard error	0.013	0.025	0.063	0.050	0.073	0.011
Adjusted R^2	.29	.26	.07	.30	.31	
No. obs.	572,089	126,649	12,050	517	22,723	734,028
Percentage	77.9	17.3	1.6	.1	3.1	100

Table 2: Item nonresponse on earnings.

	Response	Partial NR	Full NR	Total
2002 UDB, Wage and salary income				
No. obs.	81,629	8,060	772	90,461
Percentage	90.2	8.9	0.9	100
Imputation index	0	.391	1	.043
Final wage and salary income	12.823	21.138	.459	13.458
Difference final-reported income	0	9.491	.459	.850
2003 UDB, Wage and salary income				
No. obs.	89,708	1,939	580	92,227
Percentage	97.3	2.1	0.6	100
Imputation index	0	.120	1	.009
Final wage and salary income	12.836	12.514	.436	12.751
Difference final-reported income	0	1.199	.436	.028
2002 UDB, Self-employment income				
No. obs.	26,268	1,054	13,969	41,291
Percentage	63.6	2.6	33.8	100
Imputation index	0	.363	1	.348
Final self-employment income	12.846	26.139	11.271	12.652
Difference final-reported income	0	11.535	11.271	4.107
2003 UDB, Self-employment income				
No. obs.	27,636	.	13,775	41,411
Percentage	66.7	.	33.3	100
Imputation index	0	.	1	.332
Final self-employment income	12.918	.	11.216	12.352
Difference final-reported income	0	.	11.216	3.731

Table 3: Summary statistics of the distribution of earnings.

	No. obs.	Mean	Median	10th percentile	90th percentile
2002 UDB, wages and salaries					
Respondents	81,629	12.823	11.804	2.516	22.599
Partial item nonrespondents	8,060	21.138	17.753	3.654	41.354
Full item nonrespondents	772	0.459	0.104	0.010	1.080
2003 UDB, wages and salaries					
Respondents	89,708	12.836	11.808	2.693	22.487
Partial item nonrespondents	1,939	12.514	11.200	1.900	22.898
Full item nonrespondents	580	0.436	0.082	0.008	0.813
2002 UDB, self-employment income					
Respondents	26,268	12.846	8.895	0.624	26.010
Partial item nonrespondents	1,054	26.140	15.623	1.224	61.204
Full item nonrespondents	13,969	11.271	8.462	0.397	22.110
2003 UDB, self-employment income					
Respondents	27,636	12.919	8.895	0.678	26.351
Partial item nonrespondents	0	.	.	.	.
Full item nonrespondents	13,775	11.216	8.550	0.419	22.505

Table 4: Percentiles regressions of the earnings by type of nonresponse.

	Percentile	Intercept	S.E.	No. obs.
2002 UDB, wages and salaries				
Full item nonrespondents	90%	0.989	0.162	653
Full item nonrespondents	50%	0.230	0.035	653
Full item nonrespondents	10%	0.012	0.002	653
Partial item nonrespondents	90%	24.516	0.372	7575
Partial item nonrespondents	50%	16.064	0.190	7575
Partial item nonrespondents	10%	5.358	0.161	7575
Respondents	90%	24.861	0.104	75270
Respondents	50%	16.197	0.051	75270
Respondents	10%	7.171	0.068	75270
2002 UDB, self-employment income				
Full item nonrespondents	90%	27.192	0.517	13510
Full item nonrespondents	50%	11.886	0.190	13510
Full item nonrespondents	10%	0.478	0.013	13510
Partial item nonrespondents	90%	32.859	3.770	1004
Partial item nonrespondents	50%	11.786	1.022	1004
Partial item nonrespondents	10%	1.104	0.198	1004
Respondents	90%	30.993	0.503	24731
Respondents	50%	12.390	0.135	24731
Respondents	10%	1.393	0.036	24731
2003 UDB, wages and salaries				
Full item nonrespondents	90%	1.323	0.197	455
Full item nonrespondents	50%	0.173	0.012	455
Full item nonrespondents	10%	0.009	0.001	455
Partial item nonrespondents	90%	25.541	0.819	1781
Partial item nonrespondents	50%	16.127	0.431	1781
Partial item nonrespondents	10%	4.982	0.287	1781
Respondents	90%	24.810	0.102	83211
Respondents	50%	16.261	0.049	83211
Respondents	10%	7.480	0.056	83211
2003 UDB, self-employment income				
Full item nonrespondents	90%	27.717	0.634	13347
Full item nonrespondents	50%	12.103	0.186	13347
Full item nonrespondents	10%	0.531	0.014	13347
Respondents	90%	31.481	0.471	26246
Respondents	50%	12.576	0.142	26246
Respondents	10%	1.491	0.034	26246

Table 5: Summary statistics for the earnings growth rate by type of nonresponse.

	No. obs.	Mean	Median	p10	p90
2002 UDB, wages and salaries					
Respondents in t	81,629	0.965	0.000	-.356	0.579
Partial item nonrespondents in t	8,060	2.277	0.646	-.193	3.294
Fully item nonrespondent in t	772	-.234	-.743	-.995	0.656
Respondents in $t-1$	48,803	0.544	0.003	-.359	0.625
Partial item nonrespondents in $t-1$	4,383	1.637	0.333	-.307	2.766
Fully item nonrespondent in $t-1$	230	99.902	2.709	-.139	188.976
2002 UDB, self-employment income					
Respondents in t	26,268	5.245	-.024	-.575	1.298
Partial item nonrespondents in t	1,054	10.748	0.629	-.428	3.397
Fully item nonrespondent in t	13,969	4.121	0.073	-.895	6.199
Respondents in $t-1$	17,354	5.658	0.030	-.571	1.387
Partial item nonrespondents in $t-1$	616	10.997	0.313	-.511	2.915
Fully item nonrespondent in $t-1$	7,313	3.103	-.062	-.876	7.212
2003 UDB, wages and salaries					
Respondents in t	89,708	0.735	0.000	-.354	0.565
Partial item nonrespondents in t	1939	0.579	0.033	-.521	0.960
Fully item nonrespondent in t	580	-.736	-.943	-.998	-.434
Respondents in $t-1$	53,410	0.357	0.000	-.363	0.540
Partial item nonrespondents in $t-1$	1,078	0.354	-.029	-.441	1.216
Fully item nonrespondent in $t-1$	144	141.244	12.533	1.084	414.872
2003 UDB, self-employment income					
Respondents in t	27,636	2.972	-.021	-.590	1.344
Fully item nonrespondent in t	13,775	7.630	0.061	-.869	5.527
Respondents in $t-1$	17,923	4.950	0.022	-.582	1.386
Fully item nonrespondent in $t-1$	7,452	2.892	-.057	-.835	6.065

Table 6: Percentiles regressions for the earnings growth rates by type of nonresponse.

	Perc.	Intercept	S.E.	No.	Intercept	S.E.	No.
Wages and salaries							
		2002 UDB			2003 UDB		
Fully item nonresponse in $t-1$	90%	411.013	354.065	215	339.557	373.959	129
Fully item nonresponse in $t-1$	50%	10.060	1.327	215	11.482	8.718	129
Fully item nonresponse in $t-1$	10%	0.065	0.232	215	1.648	2.551	129
Fully item nonresponse in t	90%	0.223	0.714	229	-.483	2.687	118
Fully item nonresponse in t	50%	-.821	0.034	229	-.914	0.020	118
Fully item nonresponse in t	10%	-.982	0.008	229	-.997	0.003	118
Partial item nonresponse in $t-1$	90%	0.857	0.135	4098	0.555	0.260	984
Partial item nonresponse in $t-1$	50%	0.229	0.037	4098	-.039	0.021	984
Partial item nonresponse in $t-1$	10%	-.204	0.027	4098	-.347	0.061	984
Partial item nonresponse in t	90%	0.882	0.204	4076	0.480	0.223	994
Partial item nonresponse in t	50%	0.369	0.037	4076	0.025	0.015	994
Partial item nonresponse in t	10%	-.136	0.039	4076	-.393	0.062	994
Respondents in $t-1$	90%	0.316	0.020	45454	0.295	0.016	50067
Respondents in $t-1$	50%	0.003	0.002	45454	0.002	0.002	50067
Respondents in $t-1$	10%	-.225	0.008	45454	-.229	0.007	50067
Respondents in t	90%	0.279	0.018	45512	0.284	0.017	50068
Respondents in t	50%	0.000	0.002	45512	0.001	0.002	50068
Respondents in t	10%	-.227	0.007	45512	-.229	0.007	50068
Self-employment income							
		2002 UDB			2003 UDB		
Full item nonresponse in $t-1$	90%	3.215	1.354	7147	2.693	1.170	7298
Full item nonresponse in $t-1$	50%	-.044	0.008	7147	-.047	0.009	7298
Full item nonresponse in $t-1$	10%	-.792	0.022	7147	-.734	0.024	7298
Full item nonresponse in t	90%	2.788	1.131	7377	2.650	1.106	7321
Full item nonresponse in t	50%	0.079	0.009	7377	0.056	0.009	7321
Full item nonresponse in t	10%	-.834	0.022	7377	-.750	0.023	7321
Partial item nonresponse in $t-1$	90%	1.322	0.925	594	.	.	.
Partial item nonresponse in $t-1$	50%	-.002	0.115	594	.	.	.
Partial item nonresponse in $t-1$	10%	-.590	0.126	594	.	.	.
Partial item nonresponse in t	90%	2.574	0.936	602	.	.	.
Partial item nonresponse in t	50%	0.206	0.108	602	.	.	.
Partial item nonresponse in t	10%	-.508	0.123	602	.	.	.
Respondents in $t-1$	90%	1.420	0.087	16373	1.320	0.095	17049
Respondents in $t-1$	50%	0.025	0.004	16373	0.022	0.004	17049
Respondents in $t-1$	10%	-.575	0.015	16373	-.584	0.015	17049
Respondents in t	90%	1.346	0.082	16142	1.386	0.087	17026
Respondents in t	50%	-.017	0.004	16142	-.012	0.004	17026
Respondents in t	10%	-.566	0.015	16142	-.589	0.014	17026

Table 7: Worst case and reported income bounds.

Country	CI lower	LB	UBr	CI upper	UB	CI upper	No. obs.
Poverty line at 40% of median income							
Belgium	2.6	2.7	7.5	7.8	31.6	32.1	38020
Denmark	1.0	1.1	4.6	4.9	17.6	18.0	32866
France	3.0	3.2	5.2	5.4	27.9	28.2	79659
Germany-GSOEP	2.1	2.2	6.1	6.3	33.3	33.6	79034
Greece	7.2	7.4	20.3	20.6	35.0	35.4	70923
Ireland	2.5	2.7	7.6	7.8	19.2	19.6	55115
Italy	6.2	6.3	15.5	15.7	33.9	34.2	95808
Netherlands	2.4	2.5	5.9	6.1	12.0	12.2	57504
Portugal	7.1	7.2	14.0	14.2	28.0	28.3	71875
Spain	5.4	5.5	13.4	13.7	27.2	27.5	96009
UK-BHPS	6.9	7.1	11.2	11.4	15.9	16.2	57212
Poverty line at 50% of median income							
Belgium	5.9	6.1	12.2	12.6	35.0	35.5	38020
Denmark	2.3	2.5	7.0	7.3	19.0	19.4	32866
France	7.0	7.2	10.2	10.4	31.9	32.2	79659
Germany-GSOEP	4.3	4.4	9.8	10.0	35.5	35.8	79034
Greece	10.9	11.2	25.5	25.9	38.9	39.2	70923
Ireland	9.0	9.2	15.3	15.6	25.8	26.2	55115
Italy	10.1	10.3	20.7	21.0	37.9	38.2	95808
Netherlands	4.3	4.5	8.5	8.8	14.0	14.2	57504
Portugal	11.4	11.7	19.7	20.0	32.4	32.7	71875
Spain	9.0	9.1	18.4	18.6	30.8	31.1	96009
UK-BHPS	11.7	12.0	16.9	17.2	20.7	21.1	57212
Poverty line at 60% of median income							
Belgium	10.8	11.1	18.8	19.2	40.0	40.5	38020
Denmark	5.9	6.1	11.8	12.2	22.7	23.1	32866
France	12.8	13.0	17.6	17.9	37.7	38.1	79659
Germany-GSOEP	8.0	8.2	15.4	15.7	39.3	39.6	79034
Greece	15.4	15.6	31.4	31.8	43.3	43.7	70923
Ireland	15.4	15.7	23.0	23.3	32.3	32.6	55115
Italy	15.4	15.6	27.5	27.8	43.2	43.5	95808
Netherlands	8.3	8.5	13.2	13.5	17.9	18.2	57504
Portugal	17.1	17.4	26.8	27.1	38.1	38.5	71875
Spain	14.1	14.3	25.0	25.3	36.0	36.3	96009
UK-BHPS	17.5	17.9	23.5	23.8	26.6	27.0	57212

Table 8: Percentages of inconsistencies for the IV and MIV bounds.

Country	IV bounds		MIV1 bounds		MIV2 bounds		MIV3 bounds	
	standard	reported	standard	reported	standard	reported	standard	reported
Poverty line at 40% of median income								
Belgium	3.3	4.2	0.0	0.0	0.2	0.2	2.0	4.9
Denmark	1.2	1.4	5.3	5.3	1.2	2.8	15.0	17.2
France	5.4	28.6	0.0	0.0	0.3	6.8	0.2	3.2
Germany-GSOEP	2.8	15.2	0.0	0.0	0.0	0.0	0.0	0.0
Greece	4.1	13.7	0.9	1.3	0.0	0.0	1.5	2.0
Ireland	12.1	12.5	1.0	1.0	0.0	0.0	5.7	6.0
Italy	2.7	4.5	0.0	0.4	0.0	0.5	0.0	0.1
Netherlands	28.8	34.9	12.3	14.6	10.0	12.7	1.7	1.7
Portugal	7.9	10.7	0.8	5.5	0.1	0.1	10.6	13.3
Spain	0.0	0.0	9.4	11.6	0.0	0.0	0.0	0.0
UK-BHPS	43.2	50.4	0.0	0.0	6.0	9.6	0.0	0.0
Poverty line at 50% of median income								
Belgium	5.2	5.3	0.0	0.0	0.0	0.0	0.0	0.0
Denmark	2.9	4.0	3.3	4.5	4.3	5.9	11.5	17.7
France	2.1	24.9	0.0	0.0	0.3	1.7	0.3	2.1
Germany-GSOEP	4.1	12.0	0.0	1.4	0.0	0.0	0.0	0.5
Greece	6.2	14.9	0.4	1.6	0.0	0.0	0.1	0.5
Ireland	18.7	24.0	0.7	0.7	1.7	1.7	2.8	3.6
Italy	1.6	3.3	0.0	0.0	0.0	0.0	0.0	0.0
Netherlands	12.0	21.1	13.5	15.5	4.0	4.0	1.1	1.1
Portugal	6.0	13.6	0.7	2.2	0.5	0.6	18.8	23.0
Spain	0.0	0.0	8.4	10.1	0.0	1.1	0.0	0.0
UK-BHPS	45.9	53.5	0.0	0.0	5.4	6.3	0.0	0.0
Poverty line at 60% of median income								
Belgium	4.2	4.8	0.0	0.0	0.0	0.0	0.0	0.0
Denmark	5.1	5.4	1.2	1.2	0.6	1.1	10.7	15.3
France	5.5	27.5	0.0	1.2	0.0	0.0	0.3	1.5
Germany-GSOEP	1.2	23.0	0.0	2.0	0.3	1.1	0.0	0.0
Greece	1.1	3.4	0.3	0.5	0.0	0.0	0.0	0.0
Ireland	15.7	32.3	0.7	0.7	1.1	1.4	1.9	2.3
Italy	0.8	2.3	0.0	0.0	0.0	0.0	0.0	0.0
Netherlands	22.7	33.3	14.5	15.5	0.1	0.1	0.8	1.8
Portugal	11.5	14.4	0.6	0.9	0.0	0.1	8.7	12.8
Spain	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0
UK-BHPS	56.0	65.0	0.0	0.0	0.4	1.1	0.4	0.9

Table 9: Width of the IV and MIV bounds.

Country	IV bounds		MIV1 bounds		MIV2 bounds		MIV3 bounds	
	standard	reported	standard	reported	standard	reported	standard	reported
Poverty line at 40% of median income								
Belgium	26.3	3.4	27.6	4.4	26.6	4.3	28.2	4.2
Denmark	13.4	2.3	14.8	2.9	15.4	2.8	13.9	2.5
France	22.0	0.9	24.2	1.9	22.9	1.9	23.6	1.8
Germany-GSOEP	19.5	2.4	30.4	3.6	30.5	3.8	29.1	3.4
Greece	20.8	8.5	21.9	10.5	27.7	13.0	26.6	11.9
Ireland	14.5	2.9	16.2	4.8	15.4	4.6	15.9	4.4
Italy	25.1	6.0	23.8	8.6	27.1	9.0	27.1	8.8
Netherlands	7.2	1.8	8.5	2.5	8.1	2.8	8.4	3.0
Portugal	15.4	3.1	18.2	6.0	20.6	6.7	18.2	4.5
Spain	21.7	7.9	19.8	7.4	21.6	7.9	21.6	7.8
UK-BHPS	4.6	0.3	8.7	4.0	8.0	3.3	8.6	4.0
Poverty line at 50% of median income								
Belgium	25.7	3.9	28.3	6.1	27.4	5.9	28.5	5.9
Denmark	13.3	3.1	15.3	3.8	15.2	3.6	13.8	3.1
France	21.7	1.3	24.5	3.0	24.0	2.9	24.1	2.9
Germany-GSOEP	19.3	3.3	30.5	4.9	30.6	5.3	29.7	5.0
Greece	20.9	9.6	23.3	13.0	27.7	14.4	26.9	13.6
Ireland	12.8	2.1	16.3	6.0	15.5	5.6	16.2	5.8
Italy	25.2	7.4	25.8	10.0	27.2	10.4	27.3	10.2
Netherlands	6.7	1.7	8.1	2.6	8.4	3.6	8.6	3.6
Portugal	14.8	3.6	19.2	7.8	20.6	8.0	17.9	5.5
Spain	21.7	9.2	20.5	8.7	21.6	9.2	21.6	9.2
UK-BHPS	3.9	0.2	8.7	4.9	8.2	4.3	8.7	4.9
Poverty line at 60% of median income								
Belgium	25.6	5.0	28.6	7.7	28.0	7.4	28.5	7.5
Denmark	13.1	3.8	16.1	5.6	15.6	5.1	13.0	3.0
France	21.1	2.0	24.6	4.4	24.5	4.5	24.4	4.4
Germany-GSOEP	19.2	4.4	30.4	6.6	30.7	7.1	30.4	7.0
Greece	20.8	11.0	24.7	14.9	27.7	15.8	27.0	15.1
Ireland	12.3	2.9	16.5	7.2	15.8	6.9	16.4	7.1
Italy	25.2	8.8	26.4	11.7	27.3	11.9	27.4	11.8
Netherlands	6.1	1.7	7.7	2.9	8.8	4.6	8.9	4.4
Portugal	13.8	3.5	19.8	9.1	20.6	9.3	18.2	7.1
Spain	21.7	10.7	21.2	10.5	21.7	10.7	21.6	10.6
UK-BHPS	2.8	-.3	8.8	5.6	8.5	5.4	8.7	5.6

Table 10: Poverty probabilities, nonresponse probabilities and width of the worst case bounds.

Country	Poverty	Poverty for $D_i = 1$	Poverty for $D_i = 0$	$\Pr\{Y_i^r \leq c \mid D_i = 0\}$	UB-LB $\Pr\{D_i = 0\}$	UBr-LB
Poverty line at 40% of median income						
Belgium	4.1	3.8	5.2	26.9	28.9	4.8
Denmark	1.5	1.3	2.7	29.5	16.5	3.5
France	4.5	4.1	5.6	11.6	24.7	2.1
Germany-GSOEP	3.8	3.5	5.0	22.2	31.0	3.9
Greece	10.2	9.8	10.1	39.4	27.7	13.0
Ireland	3.1	3.2	2.6	32.2	16.6	4.9
Italy	8.7	8.4	9.4	28.3	27.6	9.1
Netherlands	3.1	2.9	5.7	44.4	9.4	3.4
Portugal	9.4	9.0	10.9	29.8	20.7	6.7
Spain	8.0	7.1	11.2	27.0	21.7	7.9
UK-BHPS	7.7	8.2	4.9	43.7	8.8	4.1
Poverty line at 50% of median income						
Belgium	8.4	8.5	8.4	33.3	28.9	6.1
Denmark	3.3	3.0	5.1	38.2	16.5	4.5
France	9.3	9.2	8.8	17.8	24.7	3.0
Germany-GSOEP	7.0	6.8	8.0	28.4	31.0	5.3
Greece	15.8	15.0	16.2	43.7	27.7	14.4
Ireland	10.2	11.2	5.8	37.8	16.6	6.0
Italy	13.8	13.8	13.5	33.6	27.6	10.4
Netherlands	5.5	5.3	8.6	50.4	9.4	4.0
Portugal	14.7	14.4	15.3	34.2	20.7	8.0
Spain	12.6	11.7	16.0	32.6	21.7	9.2
UK-BHPS	12.9	13.8	7.8	51.7	8.8	4.9
Poverty line at 60% of median income						
Belgium	14.8	15.3	13.1	39.8	28.9	7.7
Denmark	8.0	7.4	10.9	44.1	16.5	5.7
France	16.4	16.8	14.3	23.3	24.7	4.6
Germany-GSOEP	12.1	12.4	12.3	35.9	31.0	7.2
Greece	21.8	21.2	21.6	50.6	27.7	15.8
Ireland	17.6	19.0	11.5	44.0	16.6	7.3
Italy	20.3	20.9	18.4	37.9	27.6	11.9
Netherlands	10.2	9.8	15.0	57.2	9.4	4.7
Portugal	21.5	21.4	21.2	39.1	20.7	9.4
Spain	19.0	18.3	21.7	39.1	21.7	10.7
UK-BHPS	19.3	20.5	13.5	59.7	8.8	5.6

A Imputation of personal income components

Imputation procedure has changed across ECHP releases. In the very first release, imputation was performed by using random hotdeck imputation within classes and predictive mean matching. In the most recent releases, instead, a new imputation procedure called IVE (Imputation and Variance Estimation) has been adopted. In the following, we describe this most recent imputation procedure adopted to impute income variables when individuals return their personal questionnaire but do not answer to all income questions, i.e. when item nonresponse on personal income occurs.

When a personal income variable is missing and the person was interviewed in last wave, then the income observed in previous year is imputed to replace the missing current income. If the income in the last wave is also missing, then Eurostat uses the IVE imputation. This procedure may be viewed as a variant of the EM algorithm, because it iteratively repeats the imputation of missing values until the difference between the values obtained from two consecutive iterations is lower than a given threshold or the number of iterations exceeds a given limit.[18] In the first step of IVE, imputation is applied to variables with a low fraction of missing cases using the information from variables without missing data. In the second step, imputation is applied to variables with more severe problem of missingness, conditioning both on variables without missing data and variables imputed in the first step, and so on. The higher is the percentage of missing cases in a variable, the greater is the number of regressions to be carried out sequentially before imputing its missing values. The specific model used for the imputation depends on the type of variable to be imputed. For example, it is a linear regression model when the target variable is continuous, and a logistic regression model when the target variable is binary.[19] Imputed values of income variables are forced to lie between the minimum and the maximum values observed for respondents.

The auxiliary variables used to impute personal income amounts include the region of residence, the number of interviewed household members who work, age, gender, schooling level, marital status, the type of occupation, the employment sector and the size of the local unit in the current job, status in employment in the current and the previous job, the number of work hours and the job status. When some of these variables are missing, as it sometimes occurs, they themselves become target variables to be imputed at an earlier stage of the IVE procedure.

B Imputation for unit nonresponse within responding households

As for item nonresponse, the imputation for unit nonresponse within responding household has changed across releases.

Until the 2002 UDB, covering waves 1994–1998, the imputation was carried out by Eurostat for all countries by computing a within-household inflation factor.[20] Construction of the within-household inflation factor starts by computing a "provisional personal income" for each responding household member. This is just the sum of the different types of personal income (reported or imputed), plus the "assigned" income components (that is, the value of income components collected only at the household level divided by the number of unit respondents within the household).

The sample is then divided into 110 groups using auxiliary variables that include age classes, sex and quintiles of equivalized net monthly household income obtained from the household questionnaire. For each group g, a weighted average $\overline{Y}_g$ of provisional personal incomes is computed using the cross-sectional weights.[21] This weighted average is then assigned to each eligible household member belonging to that group, whether responding or not.

[18] The procedure has been carried out using software, developed by the Survey Research Center at the Institute for Social Research of the University of Michigan (for a description see Raghunathan, Solenberger and Hoewyk 1999).

[19] For a detailed description see Eurostat (2002).

[20] For a detailed description see Eurostat (2001).

[21] In the ECHP weights are computed to take account of the sampling design and of nonresponding households for which no questionnaire is available. See Eurostat (2002) for a description of weights computation.

Finally, the within-household nonresponse inflation factor is computed as

$$f_h = \frac{\sum_g \overline{Y}_g \sum_i 1\{i \in g\}}{\sum_g \overline{Y}_g \sum_i 1\{i \in g\}\, D_{hi}},$$

where $1\{i \in g\}$ is a 0–1 indicator equal to 1 if individual i belongs to group g, D_{hi} is a 0–1 indicator equal to 1 if individual i returns the questionnaire and 0 otherwise, and $\sum_i$ is the sum over all eligible individuals in household h. If the procedure gives as a result a value greater than 5, then the within-household nonresponse factor is set equal to missing.

In the 2003 UDB, covering waves 1994–2000, the imputation procedure adopted for unit nonresponse within responding households changed completely. The new imputation method exploits previously neglected information on individual and household income in the current and previous wave.[22] The new imputation method is applied to all countries except Finland, Ireland and the UK, that rely instead on their own imputation methods.

The new imputation procedure is no longer based on an inflation factor, but on the computation of an "additional income amount" added to the household income (after item imputation) to take account of unit nonresponse within the household.

When a household includes unit nonrespondents, total household income is computed by summing up reported and imputed personal incomes. The resulting amount is then compared with that obtained multiplying by 12 the total monthly household income reported in the last wave. The additional household income is equal to the difference between the former and the latter amount if this difference is positive, and to zero otherwise. If the household composition changes between waves or the monthly household income is missing in the last wave, then the current monthly household income is used instead. For the first wave of the panel, information on past income is not available and the values from the following waves are used.

[22] For a detailed description see Eurostat (2002).