J. R. Statist. Soc. A (2006)
169, *Part* 3, *pp.* 625–646

The effects of income imputation on microanalyses: evidence from the European Community Household Panel

Cheti Nicoletti

University of Essex, Colchester, UK

and Franco Peracchi

University of Rome "Tor Vergata", Italy

[Received June 2004. Final revision December 2005]

Summary. Social surveys are usually affected by item and unit non-response. 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 the paper is to give data users some guidelines on how to assess the effects of imputation on their microlevel analyses. We focus attention on the potential bias that is caused by imputation in the analysis of income variables, using the European Community Household Panel as an illustration.

Keywords: Income imputation; Item non-response; Panel data

1. Introduction

Social surveys are usually affected by non-response: either failed contact, or refusal to fill in the questionnaire (unit non-response) or refusal to answer specific survey questions (item non-response). If the data are missing completely at random (MCAR), i.e. the probability of non-response 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 assumption of data MCAR is far stronger than necessary, however. In practice, it can often be replaced by the weaker assumption of data missing at random (MAR). (We refer to Rubin (1976) and Little and Rubin (1987) for a formal definition of the terms MAR and MCAR.) Although this assumption allows the response probability to depend on observed variables, it imposes conditional independence between the response probability and the unobserved variables in the model of interest given the observed variables. The assumption of data MAR is important because it underlies most imputation procedures that are employed by survey methodologists to fill in the missing data.

The availability of an easy- and ready-to-use data set is clearly attractive to most applied researchers, whose main aim is typically far from understanding the response behaviour of the sample units. Unfortunately, imputation procedures may be inadequate to handle missing

Address for correspondence: Cheti Nicoletti, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK.
E-mail: nicolet@essex.ac.uk

 0964–1998/06/169625

data problems, either because they are improperly applied, or because too few variables are observed for both respondents and non-respondents that can be used to impute the missing values.

The aim of this paper is to illustrate how to evaluate the effects of imputation in microanalyses. For concreteness, we consider the effects of imputation on the analysis of income variables by using the European Community Household Panel (ECHP), a longitudinal household survey covering all countries of the European Union before the 2004 enlargement.

Much of the literature about imputation focuses on the problem of underestimation of the sampling variance of estimates that are computed by using imputed values (see for example Rubin (1989, 1996)). In this paper, we instead focus on the potential bias of estimates of a micromodel of interest. We compare various summaries of the distribution of income variables between different types of non-respondents by using the imputed values that are 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.

The remainder of the paper is organized as follows. Section 2 briefly describes the data, defines the different types of non-response affecting income variables and gives detail on the imputation procedures that are adopted in the ECHP. Section 3 describes some methods to assess consistency of imputation procedures for a model of interest and imposing the assumption of data MAR. Section 4 applies these methods to the analysis of income variables by using the ECHP. Models for household income and the earnings structure are considered in Sections 4.3 and 4.4 respectively. Finally, Section 5 offers some conclusions.

2. Non-response and imputation in the European Community Household Panel

This section briefly describes the data, defines the different types of non-response for household and personal income and gives some details on the imputation procedures that are adopted in the ECHP. (We refer to Peracchi (2002) for a more detailed description of the data.)

We use the user database (UDB) of the ECHP, which is an anonymized and easy-to-use version of the data, and focus on changes in the imputation procedures between the fourth release of the data (UDB 2002), which were issued in February 2002 and cover waves 1–5, and the fifth release (UDB 2003), which were issued in December 2003 and cover waves 1–7.

2.1. Brief description of the European Community Household Panel

The ECHP is a longitudinal survey of households and individuals, which was centrally designed and co-ordinated 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.

In its first (1994) wave, the survey covered about 60000 households and 130000 individuals in 12 countries, namely Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain and the UK. Austria, Finland and Sweden began to participate in 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 Socio-economic Panel, the Luxembourg Social Economic Panel and the British Household Panel Survey. Starting from the fourth (1997) wave, it was decided to merge the ECHP into the German Socio-economic Panel, the Luxembourg Social Economic Panel and the British

Household Panel Survey. In this paper, we focus attention on the nine countries which participated in the survey for the first five waves, namely Belgium, Denmark, France, Greece, Ireland, Italy, the Netherlands, Portugal and Spain. (Germany and the UK are excluded because the derivation of the data set, respectively from the German Socio-economic Panel and the British Household Panel Survey, has undergone changes and corrections between the two releases of the UDBs that are used in this paper. Luxembourg is excluded because the data for its harmonized national panel are not available in the 2002 release of the data.)

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

An essential feature of the ECHP is the adoption of a common questionnaire that was centrally designed by Eurostat. The questionnaire consists of a household register, mainly for record keeping and control of the sample, a household questionnaire that was submitted to a 'reference person' (usually the head of the household or the spouse or partner of the head) and a personal questionnaire that was submitted to all eligible household members. Most personal interviews were face to face and were carried out using the conventional 'paper-and-pencil' method.

2.2. Definition of income components

The information on income that is 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. (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.)

The ECHP distinguishes between six main sources of income: wages and salaries, income from self-employment or farming, pensions (old-age-related benefits and survivors' benefits), unemployment or redundancy benefits, other social benefits or grants (family-related allowances, sickness or invalidity benefits, education-related allowances, other personal benefits, social assistance and housing allowances) and non-work private income (capital income, property or rental income and private transfers received). Each income component generally consists of subcomponents, with varying level of detail. (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.) 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 or 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 non-response

We now turn to non-response to income variables among responding households, namely those where at least one eligible member returned the personal questionnaire. We do not consider non-responding households, for which no data were collected. (The ECHP takes non-responding households into account by computing weights.) Within a responding household, we allow for unit non-response, i.e. we allow for the case when eligible household members fail to return the personal questionnaire.

We also consider item non-response, 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 that are obtained by adding up different components, two types of item non-response may arise: full and partial. The former arises when all components that are needed to compute an income aggregate are missing (e.g. both the regular and the lump sum components of wage and salary earnings are missing), the latter when only some components are missing (e.g. regular wage and salary earnings are available but the lump sum component is missing).

A more common definition of item non-response refers to the case when an individual returns her personal questionnaire but fails to respond to one or more questions. In this paper we instead define item non-response in relationship to the specific income variable that is being studied. Thus, for example, item non-response to personal earnings (wages and salaries or self-employment income) occurs when the personal questionnaire is returned but some of the answers about personal earnings are missing. Considering instead household income, item non-response occurs when at least one of the household members who return their personal questionnaire fails to answer some questions on income.

Moreover, in the case of non-response to household income, we distinguish between five types of household:

(a) households with complete response (neither item nor unit non-response);
(b) households with partial item non-response (all eligible household members return their personal questionnaire but some of them fail to answer some questions on income);
(c) households with full item non-response (all eligible household members return their personal questionnaire but all of them provide no answer to questions on income);
(d) households with unit non-response (some eligible household members return the questionnaire and answer all questions on income, but some others do not return their personal questionnaire);
(e) households with both unit non-response and item non-response (only some eligible members return the questionnaire, and some of them fail to answer some questions on income).

2.4. Income imputation

In this section we briefly describe the imputation procedures and the information that is available in UDB 2002 and UDB 2003 to identify unit and item non-response.

In the case of item non-response to questions on income, Eurostat applies an imputation procedure at the individual level to replace the missing personal income components. Some

information on this procedure is provided in Appendix B. (See Eurostat (2002a) for a more detailed description.)

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

For households with unit non-response, 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, i.e.

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

where Σ_i denotes summation over all eligible members of household h, Σ_j denotes summation over all subcomponents of income, D_{hi} equals 1 if the ith individual returns the questionnaire and equals 0 otherwise, R_{hij} equals 1 if the ith individual answers the question on the jth subcomponent of personal income and equals 0 otherwise, and Y_{hij} and $\hat{Y}_{hij}$ 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 non-response. The nature of this correction has changed over time. In UDB 2002, it consists of inflating imputed household income Y_h^{I} by a 'within-household non-response inflation factor' $f_h > 1$. (The within-household non-response inflation factor is constant within households and for all subcomponents of income. For example, household income from self-employment or from wages and salaries is multiplied by the same factor, no matter whether the unit non-respondents were working as employees or self-employed in the previous year.) In UDB 2003, the correction consists instead of adding to Y_h^{I} an 'additional income amount', whose computation exploits information on income variables that were observed in the last wave. (See Appendix C for some details.)

Unfortunately, the UDB provides no flag for income imputation at the individual level. At the household level, two indicators are provided. One is the imputation ratio for item non-response, which is defined as

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

where

$$Y_h^{\mathrm{R}} = \sum_i \sum_j D_{hij}R_{hij}Y_{hij}$$

is reported household income. The other indicator is the within-household non-response inflation factor f_h for UDB 2002 and the 'additional household income' for UDB 2003. Both 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 indicators provide information that is useful to distinguish between the four types of income non-response at the household level that were introduced in Section 2.3.

3. Assessing imputation under data missing at random

This section gives conditions under which we can consistently estimate the parameters of a model of interest either by dropping the missing data or by replacing the missing data with their imputations. It also proposes an informal method for checking whether the imputation is 'congenial' (Meng, 1994) under the maintained assumption of data MAR. (Imputations usually impose a data MAR assumption, and the ECHP is no exception.)

Our aim is not to give guidelines to improve imputation procedures, but rather to suggest to applied researchers how to assess whether the imputed data that are available in most public data sets can be used to estimate the parameters of interest consistently. For consistent estimation of an estimator's variance, we refer to Rubin (1996), Fay (1996) and Robins and Wang (2000), where multiple-imputation procedures are considered to overcome the problem of inconsistent estimation of an estimator's variance when a single imputation is used.

3.1. Model estimation in the presence of missing outcome data

Suppose that we want to estimate a statistical model involving a p-dimensional population parameter θ which is implicitly defined through the conditional moment restriction

$$E[\psi(X, Y; \theta)|X] = 0, \tag{1}$$

where Y is the outcome of interest, X is a vector of $r \geqslant p$ explanatory variables and ψ is some (real-valued) moment function. This conditional moment restriction implies the unconditional moment restriction $E[\psi(X, Y; \theta)X] = 0$, which in turns provides a basis for estimating the population parameter θ by the generalized method of moments (Hansen, 1982).

In particular, if $r = p$ and $\{(X_i, Y_i)\}_{s=1}^n$ is a random sample from (X, Y), then θ may be estimated consistently by a root of the estimating equation

$$n^{-1} \sum_{i=1}^{n} \psi(X_i, Y_i; \theta) X_i = 0. \tag{2}$$

Many common estimators, such as the least squares and quantile regression estimators that are discussed in Section 4.1, may be interpreted in this way. For simplicity, we shall only consider the case of simple random sampling, although the extension to other types of sampling scheme is straightforward. (If the sampling scheme is more complex than simple random sampling, then the relevant conditional moment restriction becomes $E[\psi(X, Y; \theta)SW|X] = 0$, where S is a binary variable indicating whether a population unit is selected into the sample and W is a weight that is inversely proportional to the conditional probability of sample inclusion. See Wooldridge (1999, 2001) for details.)

What happens if the data on the outcome variable Y are partly missing, e.g. because of non-response? What if the missing data on Y are replaced by imputed values?

Let D_i be a binary random variable equal to 0 if Y is missing for the ith unit and equal to 1 otherwise. A truncated data estimator of θ is a root $\hat{\theta}_{\mathrm{T}}$ of

$$n^{-1} \sum_{i=1}^{n} \psi(X_i, Y_i; \theta) X_i D_i = 0.$$

This estimator only uses the subsample with complete observations on Y. Consistency of $\hat{\theta}_{\mathrm{T}}$ requires that

$$E[\psi(X, Y; \theta)D|X] = 0. \tag{3}$$

If D does not depend on observed or unobserved variables, then we say that the data are MCAR. In this case, condition (1) still holds for the subsample of units with $D_i = 1$ and so a truncated data estimator is consistent for θ.

In practice, the condition of data MCAR is rarely satisfied as D typically depends on observed variables, and sometimes even on unobserved variables such as Y. When D depends only on observed variables, we say that the data are MAR. The condition of data MAR implies independence between D and Y given a set of observed variables. Note that the content of the condition of data MAR depends on the conditioning variables. For example, D and Y may be independent given the explanatory variables in the model of interest, X. Alternatively, D and Y may be independent only after conditioning on a larger set of variables, say (X, X^+), where X^+ is a vector of additional variables that are relevant in explaining the probability of observing Y.

We show in Appendix A, proof 1, that the conditional moment restriction (3) holds under any of the following three assumptions:

(a) data MCAR, or $D \perp\!\!\!\perp X, Y$,
(b) independence between D and Y given X, or $D \perp\!\!\!\perp Y|X$, or
(c) independence between D and Y given (X, X^+) and independence between Y and X^+ given X, or $D \perp\!\!\!\perp Y|(X, X^+)$ and $Y \perp\!\!\!\perp X^+|X$,

where the symbol '$\perp\!\!\!\perp$' means independence. The second part of assumption (c), namely $Y \perp\!\!\!\perp X^+|X$, is equivalent to an instrumental variable restriction, as X^+ consists of variables that are potentially relevant in explaining the response probability but irrelevant for the model of interest. Examples of such variables are the characteristics of the data collection process, which are generally assumed to be irrelevant by empirical researchers and survey methodologists for respectively the model of interest and the imputation model. (As a matter of fact, the imputation procedure that is adopted in the ECHP does not use any of these data collection characteristics to impute the missing variables.)

Now let Y_i^* denote the imputed value of Y for the ith unit. An imputed data estimator of θ is a root $\hat{\theta}_I$ of

$$n^{-1} \sum_{i=1}^{n} \{\psi(X_i, Y_i; \theta) D_i + \psi(X_i, Y_i^*; \theta)(1 - D_i)\} X_i = 0.$$

Consistency of $\hat{\theta}_I$ requires that

$$E[\psi(X, Y; \theta) D + \psi(X, Y^*; \theta)(1 - D)|X] = 0, \tag{4}$$

where missing Y has been replaced by its imputation Y^*.

We assume that the imputation model for Y uses a set Z of auxiliary variables which are observed for all units, and is such that

$$E[\psi(X, Y^*; \theta)|Z, D = 0] = E[\psi(X, Y; \theta)|Z]. \tag{5}$$

Then four different cases are possible:

(i) $Z = X$,
(ii) X is a subset of the variables in Z,
(iii) Z is a subset of the variables in X and
(iv) the set of variables X neither is included nor includes Z.

In case (i), the imputed data estimator $\hat{\theta}_I$ is consistent for θ under any of the three assumptions (a)–(c), as shown in Appendix A, proof 2.

In case (ii), let $Z = (Z_1, Z_2)$, where $Z_1 = X$ and Z_2 are auxiliary variables that are considered in the imputation procedure but excluded from the model of interest. The imputed data estimator is consistent for θ under any of the following assumptions (see Appendix A, proof 3):

(d) independence between D and Y given Z, or $D \perp\!\!\!\perp Y|Z$,
(e) independence between D and Y given (X^+, Z) and independence between Y and X^+ given Z, or $D \perp\!\!\!\perp Y|(X^+, Z)$ and $Y \perp\!\!\!\perp X^+|Z$.

The second condition in (e) is an instrumental variables assumption which requires the variables in X^+ to be relevant for the response probability model but irrelevant for both the model of interest and the imputation model.

If the variables in Z_2 are important in explaining both Y and D, then assumption (d) is more plausible than assumption (b), whereas assumption (e) is more plausible than assumption (c). In this case, it is better to use the imputed data estimator than the truncated data estimator. If the auxiliary variables Z_2 are instead relevant only for the response probability model (i.e. $Y \perp\!\!\!\perp Z_2|X$), then $D \perp\!\!\!\perp Y|Z$ implies that the truncated data estimator is also consistent. If $Y \perp\!\!\!\perp (Z_2, X^+)|X$, then $D \perp\!\!\!\perp Y|(X^+, Z)$ and so the truncated data estimator and the imputed data estimator are both consistent. (The proof of these two results is essentially the same as part (c) of proof 1 in Appendix A, after conditioning and marginalizing with respect to Z (first result) or with respect to Z_2 and X^+ (second result) rather than with respect to X^+ only.)

Finally, in case (iii), the moment condition (4) does not hold even if (d) or (e) applies, whereas, in case (iv), the moment condition (4) does not hold, neither if $D \perp\!\!\!\perp Y|Z$ nor if $D \perp\!\!\!\perp Y|(X, Z)$. Applied researchers should therefore be careful with using imputed data whenever the model of interest includes explanatory variables that have not been used in the imputation.

3.2. When is imputation inadequate?

In general, survey statisticians try to include in the imputation all the relevant explanatory variables. Schafer (1997) suggested that all variables that are relevant to explain either Y or D should be considered. (However, under the assumption $D \perp\!\!\!\perp Y|Z$, or under the double assumption $D \perp\!\!\!\perp Y|X^+, Z$ and $Y \perp\!\!\!\perp X^+|Z$, the use of the variables X^+ which are relevant for the response probability but irrelevant for the variable Y is not necessary.) Unfortunately, considering all the variables that are relevant for Y or D may not be practical, as multicollinearity or degree-of-freedom problems may lead to difficulties in identifying the coefficients that are associated with a too large set of explanatory variables. This means that imputation procedures must impose exclusion restrictions which may be at odds with the restrictions that are imposed by applied researchers in their model of interest.

As a general guideline for applied researchers, we therefore suggest checking whether the model of interest includes any relevant variable which has been omitted from the imputation procedure. If all explanatory variables are used as auxiliary variables in the imputation, then we suggest using the imputed data and possibly correcting the estimator's variance by using multiple-imputation methods.

If the model of interest involves a non-linear transformation of the outcome variable Y, say $g(Y)$, then an imputation such that $E[Y^*|Z, D=0] = E[Y|Z]$ does not imply that $E[g(Y)|Z] = E[g(Y^*)|Z, D=0]$. This is because $g(E[Y|Z]) \neq E[g(Y)|Z]$ if the function $g(\cdot)$ is not linear. So, for example, if the imputation model for Y is a linear regression, then considering a quantile regression model for Y or a model involving a monotone transformation of Y, such as the log-transformation or a categorization of Y (e.g. a dummy variable indicating whether Y is below a given threshold), may lead to inconsistent estimates. More generally, if the imputation model

is a linear regression model, then an imputed data estimator based on a moment function $\psi(\cdot)$ that is non-linear in Y may be inconsistent.

To summarize, there are two main reasons why using imputed data may lead to invalid inference:

(a) the model of interest contains explanatory variables that have been omitted from the imputation model;
(b) the model of interest contains non-linear transformations of the outcome variable.

More generally, following Meng (1994), we say that the model of interest and the imputation models are uncongenial when they are based on different parametric assumptions or on different sets of explanatory variables. Nevertheless, as suggested by Meng (1994), even when the models are uncongenial, using the imputed data may produce parameter estimates that are not significantly different from those that would be obtained by using a congenial imputation procedure. To verify this, Meng (1994) suggested using importance weights. Unfortunately, this requires knowing the exact imputation model that is used to fill in the missing data, which is generally not possible, and is definitely not possible by using the UDB of the ECHP.

Therefore, we suggest a different method for assessing imputation procedures that omit some of the explanatory variables that are considered in the model of interest (cases (iii) and (iv) above) or consider a different specification for the imputation model. We assume that the model of interest is correctly specified; more specifically, we assume that the moment condition (1) holds when using a random sample, meaning in particular that $Y \perp\!\!\!\perp Z|X$. (If this is not true, then it is obviously necessary to include the omitted variables.) Moreover, we assume that either assumption (b) or assumption (c) in Section 3.1 holds. Since we assume that all observed variables that are relevant to explain Y have been considered in the model of interest, the variables that are considered in the imputation model but omitted from the model of interest should only be relevant for the response probability model. Therefore, these omitted variables play the same role of the variables X^+ in condition (c).

Under the above assumptions, the following two conditional moment restrictions hold:

$$E[\psi(X, Y; \theta) D|X] = 0, \tag{6}$$

$$E[\psi(X, Y; \theta)(1 - D)|X] = 0. \tag{7}$$

Therefore, estimating two separate models for the respondents and the non-respondents should produce results that are not significantly different. By replacing the missing Y with the imputed Y^*, we can therefore assess whether estimations that are based on the moment restrictions (6) and (7) produce similar results. If the two sets of estimated parameters are not significantly different, then we cannot reject consistency of the imputation. If equality of the parameters is instead rejected, then the imputation model is inadequate.

Under cases (iii) and (iv), the vector of auxiliary variables Z does not include all explanatory variables X that are considered in the model of interest. If these excluded variables are relevant to explain Y, then the fit should be better for the regression of Y on X for $D = 1$ than for the regression of Y^* on X for $D = 0$. A possible measure of goodness of fit is the adjusted R^2 for linear regression, or the pseudo-R^2 for linear quantile regression or generalized linear models. A higher value of R^2 for the regression of Y on X for respondents would imply that the imputation procedure is not adequate for the model of interest.

Under cases (i) and (ii) we would instead expect a higher R^2 when estimating the regression model for non-respondents by using imputed data, at least if the model of interest and the imputation model are not very different. In the extreme and rare situation where the imputation

model and the model of interest are identical, then we would expect an R^2 equal to 1 when estimating the regression model of interest for non-respondents by using deterministically imputed data. In conclusion, when the R^2 for the respondents is higher than for the non-respondents, then the adequateness of the imputation method is doubtful. Vice versa, when the R^2 for the respondents is lower than for the non-respondents, we can only infer that the imputation model did not omit relevant explanatory variables that are included in the model of interest.

4. Empirical results

We now apply the methods that were described above to assess the effects of imputation on estimates of statistical models for household income and personal earnings. After describing the models in Section 4.1 and defining the explanatory variables in Section 4.2, Section 4.3 focuses on the effects of imputation on household income, whereas Section 4.4 focuses on personal earnings. We confine attention to static models. Considering dynamic models goes beyond the aim of this paper, as it requires an assessment of the selection effects of panel attrition.

4.1. The models of interest

We estimate models for both the mean and the selected quantiles of the logarithm of household income and personal earnings. (By household income we mean equivalized household income, defined as the ratio of total household income and the modified Organisation for Economic Co-operation and Development equivalence scale.) We consider both marginal and conditional models to assess to what extent the possible selection bias is due to observed variables. We also check whether the model of interest and the imputation model are uncongenial because they are based on different parametric assumptions.

Let Y be the outcome of interest (specifically, log(household income) or log(personal earnings)) and let X be a set of explanatory variables. To describe the relationship between Y and X we employ both mean and quantile regression models. (We refer to Koenker and Bassett (1978), Buchinsky (1998), Koenker and Hallock (2001) and Koenker (2005) for an introduction and a review of recent advances in quantile regression.) A linear mean regression model assumes that

$$E[Y|X] = X^{\mathrm{T}}\theta, \tag{8}$$

where θ is a vector of coefficients. The conditional moment restriction that is implied by this model is

$$E[Y - X^{\mathrm{T}}\theta|X] = 0. \tag{9}$$

A linear qth quantile regression model assumes instead that

$$\mathrm{quant}_q(Y|X) = \inf\{y : F(y|X) \geqslant q\} = X^{\mathrm{T}}\theta_q, \qquad 0 < q < 1,$$

where quant_q and F respectively denote the qth conditional quantile and the conditional distribution function of Y given X and θ_q is a vector of coefficients. The conditional moment restriction that is implied by this model is

$$E[q - I(Y \geqslant X^{\mathrm{T}}\theta_q)|X] = 0,$$

where $I(A)$ equals 1 if event A occurs and equals 0 otherwise.

Quantile regression has three main advantages over mean regression:

(a) it helps to characterize better the relationship between X and Y by allowing the effect of X on Y to be different at different quantiles of the conditional distribution of Y;

(b) it produces estimates that are equivariant under monotone transformations of Y;
(c) quantile regression estimators have better robustness properties than least squares.

Because of the equivariance property, we can estimate a conditional quantile of the logarithm of Y by simply applying the log-transformation to the corresponding estimated quantile of Y. The ECHP imputation procedures for individual income components are based on linear mean regression models for the logarithm of income. Unfortunately, the equivariance property does not hold for mean regression, which can lead to inconsistent estimation of models of interest when the outcome variable differs from that used in the imputation model even by a monotone transformation.

4.2. Explanatory variables

The explanatory variables for household income and personal earnings are selected following the standard practice in the economics literature (see, for example, Mincer (1974), Heckman *et al*. (2003) and Cappellari and Jenkins (2004)). Thus, our models for log(household income) are linear in the following variables: the age of the reference person (the person who fills in the household questionnaire, usually the head of the household or the spouse or partner of the head), the size of the household, the number of children who are aged less than 16 years, the number of workers in the household and dummy variables for schooling attainments and marital status of the reference person. In both UDB 2002 and UDB 2003, the imputation for unit non-response within responding households uses as auxiliary variable monthly household income, either in the last or in the current wave, which under some assumptions may be a good proxy for the missing household income. For obvious reasons, we do not use this auxiliary variable as explanatory variable in our model of interest.

Our models for log(personal earnings) contain instead the following explanatory variables: experience (a proxy for work experience, defined as the difference between the current age and the age at which the person started her or his working life), its square, the number of children and indicators for gender, schooling attainments and not having a spouse. To these variables we add a subset of the auxiliary variables that are used in the imputation. (We do not consider the effect of imputation for unit non-response on personal earnings because this imputation is performed in the ECHP at the household level.) The full list of auxiliary variables that are considered in the imputation are the number of workers in the household, age, gender, level of schooling, region of residence, occupation in the current job, status in employment, job status, total number of hours worked per week and main activity and size of the local unit where the person is working. Note that the imputation procedure for item non-response does not include as auxiliary variables the dummy for not having a spouse, the number of children and experience, which are instead considered as important explanatory variables in the economics literature.

Also note that, when earnings are missing, a few auxiliary variables are also likely to be missing. (The missing auxiliary variables in the ECHP are imputed before imputing the income variables, but they are left missing in the UDB.) In the case of full item non-response on earnings, the only auxiliary variables with a percentage of missing that is lower than 50% (excluding the variables that have already been included in the above earnings model) are the number of workers in the household, age and region. These auxiliary variables are the only additional variables that we consider in our earnings models.

In all the models considered, we transform the continuous explanatory variables into deviations from their mean. Further, when using categorical indicators, the category omitted is always the most frequent. In this way, the intercept of mean regression models represents the

mean of the outcome variable (log(household income) or log(personal earnings)) for a 'base-line' unit (a household or individual), namely one with continuous explanatory variables equal to their average value and categorical explanatory variables equal to the reference category. The intercept of quantile regression models represents instead the quantile of the outcome variable for the base-line unit.

Since regression models are defined for the logarithm of income variables, it would be useful to transform the estimated mean and quantiles of log-income for the base-line unit into the estimated mean and quantiles of income. For quantile regressions, it is enough to take the exponential of the intercept. This simple transformation produces instead downward biased estimates in the case of mean regressions, because mean regression estimates are not equivariant under monotone transformations. Thus, in our tables, we report the exponential of the intercept for quantile regressions and the 'smearing estimator' that was proposed by Duan (1983) for mean regression.

4.3. Household income

Household income is partially missing whenever some eligible household members do not answer some questions on income (item non-response) or they do not return their personal questionnaire (unit non-response). The ECHP adopts different imputation procedures in the two cases. It is therefore important to distinguish between households with only item non-response, households with only unit non-response and households with both item and unit non-response.

Individuals who return their personal questionnaire but do not answer any question on income are also very likely to fail to answer some other questions. Imputation of personal income variables for this type of individuals is especially challenging, as it requires first to impute the potentially missing auxiliary variables. For this reason, we keep households with full item non-response separate in assessing imputation.

We are interested in checking whether there are systematic differences in the distribution of household income across the different types of responding households. More specifically, we are especially interested in assessing imputation for unit non-response within responding households, which is carried out at household level. Assessing imputation for item non-response at the personal level is instead the main focus of the next section, where personal earnings are considered. For the analysis of household income we exclude Ireland, which provided its own imputation procedure for UDB 2003. For the other countries that are considered in our sample, the changes between UDB 2002 and UDB 2003 arise from changes in the imputation procedure that was adopted for unit non-responses and to possible corrections of errors in the data that occurred between the two releases.

Table 1 shows the importance of item and unit non-response for the five types of responding households that were defined in Section 2.3. We use three different concepts of household income, namely reported income Y_h^{R} (the sum of the personal incomes that are 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 adjusted by adding the additional household income). (All household income variables have been normalized by dividing them by the modified Organisation for Economic Co-operation and Development equivalence scale.) Table 1 also shows the item imputation ratio defined as the ratio $Y_h^{\mathrm{R}}/Y_h^{\mathrm{I}}$ between the reported and the imputed household income (and therefore equal to $1 - W_h$), the unit imputation ratio defined as the ratio $Y_h^{\mathrm{I}}/Y_h^{\mathrm{F}}$ between the imputed and the final household income and the total imputation ratio defined as the ratio $Y_h^{\mathrm{R}}/Y_h^{\mathrm{F}}$ between the reported and the final household income (and therefore equal to the product of the previous two ratios). All three ratios vary between 0 and 1, with 0 corresponding to cases when the entire

Table 1. Average values of imputation ratios and of household income by type of non-response

	Results for the following types of response or non-response:					*Total*
	Complete response	*Type of non-response*				
		Only item	*Full item*	*Only unit*	*Item/unit*	
UDB 2002						
Number of observations	464147	99281	11427	518	22578	597951
%	77.6	16.6	1.9	0.1	3.8	100.0
Item imputation ratio	1.000	0.769	0.000	1.000	0.658	0.930
Unit imputation ratio	1.000	1.000	1.000	0.718	0.692	0.988
Total imputation ratio	1.000	0.769	0.000	0.718	0.490	0.923
Reported income	8.681	7.972	0.000	6.316	4.704	8.245
Item imputed income	8.681	10.402	6.318	6.316	6.895	8.852
Final income	8.681	10.402	6.318	8.783	10.467	8.989
UDB 2003						
Number of observations	465360	104993	11857	377	15192	597779
%	77.8	17.6	2.0	0.1	2.5	100.0
Item imputation ratio	1.000	0.771	0.000	1.000	0.611	0.930
Unit imputation ratio	1.000	1.000	1.000	0.673	0.637	0.990
Total imputation ratio	1.000	0.771	0.000	0.673	0.445	0.926
Reported income	8.682	7.827	0.000	5.915	4.096	8.242
Item imputed income	8.682	10.035	6.387	5.915	5.926	8.803
Final income	8.682	10.035	6.387	9.621	9.395	8.893

household income comes from imputation (item, unit or total) and 1 corresponding to cases when no imputation takes place. The columns of Table 1 correspond to the different types of responding household, whereas the rows correspond to the variables (imputation ratios and household incomes) for which the average is computed. Table 1 is divided into two parts: the top part shows the results by using UDB 2002, whereas the bottom part shows the results for UDB 2003.

Item imputation is more important than unit imputation. About 22% of the households have problems of item non-response for their members, but only 3–4% of them have problems of unit non-response. The last column of Table 1 shows that the average value of the item imputation ratio (0.930 for both the 2002 and the 2003 UDB) is smaller than that of the unit imputation ratio (0.988 and 0.990 respectively for the 2002 UDB and the 2003 UDB). This implies that, on average, reported household income is inflated by 1.2–1.0% because of unit imputation, and by 7.5% because of item imputation.

In the case of full item non-response, the imputation procedure in the ECHP delivers a lower average household income for non-respondents than for respondents. (Breaking down Table 1 by country, we again find that the average value of final household income for full item non-responding households is lower than for fully responding households.)

To check whether this difference depends on the different characteristics of these households, we estimate mean and median regression models whose covariates consist of the age of the reference person, the size of the household, the number of children who are aged less than 16 years, the number of workers, a dummy variable for a reference person without a spouse and dummy variables for the schooling attainments of the reference person. The intercepts of these models represent the mean or median of log-income for a base-line household, namely a

Table 2. Mean and median regression models for log(household income) by type of non-response

	Results for the following types of response or non-response:				
	Complete response	*Type of non-response*			
		Only item	*Full item*	*Only unit*	*Item/unit*
2002 UDB					
Number of observations	448290	95849	11291	513	22156
Mean regression					
Transformed intercept	10.060	11.743	8.929	10.075	11.847
Standard error	0.018	0.053	0.298	0.785	0.124
R^2	0.274	0.259	0.088	0.121	0.214
Median regression					
Transformed intercept	8.968	10.145	5.707	9.130	10.231
Standard error	0.014	0.045	0.186	0.802	0.125
R^2	0.206	0.178	0.050	0.121	0.168
2003 UDB					
Number of observations	462114	104025	11752	377	15008
Mean regression					
Transformed intercept	10.083	11.466	8.522	11.910	10.952
Standard error	0.020	0.046	0.300	0.657	0.104
R^2	0.279	0.266	0.078	0.359	0.326
Median regression					
Transformed intercept	8.991	10.137	6.015	9.915	10.011
Standard error	0.013	0.047	0.133	0.749	0.130
R^2	0.208	0.182	0.047	0.205	0.194

household whose size, number of workers and number of children are equal to the average and whose reference person has age that is equal to the average, is married and completed secondary education.

To facilitate comparisons, Table 2 presents the estimates of median and mean income of the base-line household, denoted by 'Transformed intercept', for each type of responding households. We compute these estimates by using the smearing estimator for mean regression and the exponential of the intercept for median regressions. Table 2 also reports the standard errors of these estimates and the R^2 of each regression. (The standard errors have been estimated by using the bootstrap. By R^2 we mean the adjusted R^2 for mean regression and the pseudo-R^2 for median regression.)

The potential bias of the imputation procedure for household income does not seem to disappear after controlling for the characteristics of the households. The intercepts are indeed very low for households with full item non-response. (We also consider mean regressions with weights provided in the UDBs to take into account the sampling design and the presence of non-responding households. The results are qualitatively the same in this case.)

To check 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. In UDB 2002, a large fall in the R^2 is observed for households with unit non-response, full income non-response and

both unit and item non-response relatively to those with only item non-response. In UDB 2003, a fall in the R^2 is instead observed only for full item non-responding households. Looking at the regressions for households with only unit non-response, we find a considerable increase in the R^2 in UDB 2003 relatively to UDB 2002. Moreover, whereas in UDB 2002 the estimated coefficients differ substantially between respondents and non-respondents, in UDB 2003 they tend to be much more similar. This evidence suggests that the variables that are omitted from the unit imputation procedure in UDB 2002, but considered in our mean and median regresssions, are better taken into account by the unit imputation procedure in UDB 2003. In other words the unit imputation model in UDB 2003 more closely resembles the mean and regression models that we use.

A similar result can also be observed for households with both unit and item non-respondents, with the R^2 increasing from 0.214 in UDB 2002 to 0.326 in UDB 2003 for mean regression, and from 0.168 in UDB 2002 to 0.194 in UDB 2003 for median regression.

For full income non-responding households, the relationship between household income and the explanatory variables is not very strong. Moreover, the estimated coefficients are quite different from those of responding households. This reflects problems with the imputation procedure, which seems to underestimate household income for full income non-responding households. Income underestimation for fully item non-responding households is as serious in UDB 2003 as it was in UDB 2002.

4.4. 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 by using purchasing power parities.

To examine imputation at the individual level, we focus on households with a unique earner of the specific income that is considered. This allows us to use the imputation ratio, computed at the household level, as an individual imputation ratio taking value 1 in full item non-response, value 0 in the opposite case of full response and values between 0 and 1 in cases of partial item non-response. We do not consider households with both item and unit non-response, and we focus attention only on non-response to earnings. (Item non-response for wages and salaries (or self-employment income) refers here to cases where an individual returns her personal questionnaires but does not answer all the income questions that are needed to compute her personal wages and salaries (or self-employment income).) By reported and final earnings we mean respectively personal earnings before and after item imputation.

Table 3 shows the importance of imputation on earnings (wages and salaries and self-employment income), as measured by the percentage of imputed units and the average imputation ratio. It also shows the mean of final earnings, the mean difference between final and reported earnings and the median, the 10th percentile P10 and the 90th percentile P90 of the final earnings. (We use reported earnings for the respondents and imputed earnings for the non-respondents.)

Results are presented separately for respondents, partial item non-respondents, full item non-respondents and the full sample. There is a strong association between the type of income and the nature and importance of item non-response. The percentage of non-respondents is much higher for self-employment income than for wages and salaries. Wages and salaries are mainly affected by partial item non-response, whereas self-employment income is affected only by full item non-response. Going from UDB 2002 to UDB 2003, the number of respondents changes slightly, owing to minor corrections between the two UDB releases.

Table 3. Item non-response on earnings

	Respondents	*Partial item non-respondents*	*Full item non-respondents*	*Total*
UDB 2002, wage and salary income				
Number of observations	67812	1393	554	69759
%	97.2	2.0	0.8	100.0
Imputation ratio	0.000	0.097	1.000	0.010
Final wage and salary income	12.529	13.285	0.233	12.446
Difference final – reported income	0.000	1.217	0.233	0.026
Median	11.549	12.457	0.042	11.498
P10	2.527	3.196	0.008	2.338
P90	21.933	22.882	0.565	21.802
UDB 2003, wage and salary income				
Number of observations	67296	1317	543	69156
%	97.3	1.9	0.8	100.0
Imputation ratio	0.000	0.091	1.000	0.010
Final wage and salary income	12.663	13.693	0.446	12.587
Difference final – reported income	0.000	1.173	0.446	0.026
Median	11.634	12.394	0.078	11.595
P10	2.685	3.310	0.008	2.514
P90	22.066	23.712	0.829	22.051
UDB 2002, self-employment income				
Number of observations	21936	—	13207	35143
%	62.4	—	37.6	100.0
Imputation ratio	0.000	—	1.000	0.376
Final self-employment income	12.493	—	10.867	11.882
Difference final – reported income	0.000	—	10.867	4.084
Median	8.354	—	8.366	8.362
P10	0.556	—	0.400	0.465
P90	25.421	—	21.400	23.775
UDB 2003, self-employment income				
Number of observations	21853	—	13383	35236
%	62.0	—	38.0	100.0
Imputation ratio	0.000	—	1.000	0.380
Final self-employment income	12.651	—	11.263	12.124
Difference final – reported income	0.000	—	11.263	4.278
Median	8.449	—	8.588	8.513
P10	0.597	—	0.424	0.493
P90	25.626	—	22.520	24.193

Tables 4 and 5 give instead summaries of four different regressions: the mean, the median, the 10th percentile P10 and the 90th percentile P90 regressions of log-earnings on experience, its square and indicators for people without a spouse, schooling, sex, number of children, number of workers in the household, age and age squared. For each regression, we report the transformed intercept (Duan's smearing estimate for the mean regressions and the exponential transformation for percentile regressions), its estimated standard error SE, the regression R^2 and the number of observations. (As before, the standard errors of the transformed intercepts have been estimated by using the bootstrap, whereas the R^2 is the adjusted R^2 for mean regression and the pseudo-R^2 for quantile regression.) The transformed intercepts provide estimates of the mean, the median, the 10th and the 90th percentile of earnings for a base-line individual, namely a married man with secondary education completed and with age, experience,

Table 4. Mean and percentile regressions of wages and salaries by type of non-response

	Regression	*Transformed intercept*	*SE*	R^2	*Number of observations*
2002 UDB					
Respondents	Mean	17.917	0.105	0.204	61407
Partial item non-respondents	Mean	18.827	0.806	0.190	1270
Full item non-respondents	Mean	0.362	0.060	0.217	378
Respondents	Median	15.073	0.067	0.120	61407
Partial item non-respondents	Median	16.143	0.608	0.110	1270
Full item non-respondents	Median	0.157	0.034	0.186	378
Respondents	P10	6.997	0.109	0.153	61407
Partial item non-respondents	P10	7.623	0.703	0.149	1270
Full item non-respondents	P10	0.014	0.002	0.073	378
Respondents	P90	25.896	0.190	0.126	61407
Partial item non-respondents	P90	25.703	1.021	0.166	1270
Full item non-respondents	P90	0.674	0.117	0.077	378
2003 UDB					
Respondents	Mean	20.476	0.108	0.262	62465
Partial item non-respondents	Mean	21.547	1.072	0.225	1218
Full item non-respondents	Mean	0.728	0.108	0.176	423
Respondents	Median	17.416	0.086	0.167	62465
Partial item non-respondents	Median	18.622	0.576	0.156	1218
Full item non-respondents	Median	0.189	0.043	0.169	423
Respondents	P10	9.959	0.146	0.178	62465
Partial item non-respondents	P10	9.704	1.086	0.158	1218
Full item non-respondents	P10	0.013	0.004	0.047	423
Respondents	P90	25.631	0.166	0.187	62465
Partial item non-respondents	P90	26.151	1.351	0.229	1218
Full item non-respondents	P90	2.295	0.721	0.100	423

Table 5. Mean and percentile regressions of self-employment income by type of non-response

	Regression	*Transformed intercept*	*SE*	R^2	*Number of observations*
2002 UDB					
Respondents	Mean	15.903	0.258	0.093	20435
Full item non-respondents	Mean	14.648	0.325	0.069	12785
Respondents	Median	10.614	0.128	0.054	20435
Full item non-respondents	Median	10.809	0.168	0.033	12785
Respondents	P10	0.860	0.047	0.037	20435
Full item non-respondents	P10	0.473	0.013	0.010	12785
Respondents	P90	28.299	0.563	0.035	20435
Full item non-respondents	P90	23.225	0.424	0.026	12785
2003 UDB					
Respondents	Mean	19.769	0.399	0.106	20701
Full item non-respondents	Mean	20.074	0.567	0.075	12983
Respondents	Median	13.291	0.187	0.065	20701
Full item non-respondents	Median	13.550	0.313	0.036	12983
Respondents	P10	1.571	0.102	0.044	20701
Full item non-respondents	P10	0.581	0.027	0.011	12983
Respondents	P90	31.838	0.760	0.060	20701
Full item non-respondents	P90	29.295	1.031	0.036	12983

number of children and number of workers in the household equal to average values in the sample.

Table 3 shows that averages and percentiles of wages and salaries are similar for partial item non-respondents and respondents but are much lower for full item non-respondents. These differences persist even after controlling for a set of explanatory variables (Table 4), which suggests underestimation of wages and salaries income for full item non-respondents.

To check whether the imputed data for full item non-respondents cannot reproduce the relationship between wages and salary income and the set of explanatory variables that were defined above, we assess the presence of big differences in the R^2 between the respondents and the full item non-respondents. Median and mean regressions have in general higher R^2 for the full item non-respondents than for the respondents whereas the opposite is true for the 10th and 90th percentile regressions (Table 4). This could indicate that the imputation model for full item non-respondents uses all relevant explanatory variables but is incongruent with percentiles regression models, especially if extreme percentiles are used (10th and 90th percentiles).

Turning to self-employment income, Table 3 reveals only small differences in the sample statistics between respondents and full item non-respondents. Self-employment income seems slightly underestimated for full item non-respondents, especially when considering the 10th and 90th percentiles. The evidence of underestimation disappears at the mean and the median after we control for a set of explanatory variables (Table 5) but persists at the 10th and 90th percentile. The regressions' R^2 are always lower for full item non-respondents than for respondents. This may indicate omission of relevant auxiliary variables in the imputation model but may also (and perhaps more plausibly) indicate lack of information on auxiliary variables for full item non-respondents to self-employment income. (Whereas work experience and its square are not used as auxiliary variables in the imputation model for personal earnings, we use them as explanatory variables in our earninigs equations.)

Thus, it seems that the imputation procedure that was adopted to solve the full non-response problem produces seriously underestimated values for wages and salaries of full item non-respondents. However, because the percentage of full item non-respondents is quite low (0.8%; see Table 3), the bias in the average wage and salary computed by using all individuals is likely to be small. However, although full item non-response for self-employment earnings is high (more than 37%), the conditional and unconditional mean and median of self-employment income do not differ significantly for respondents and full item non-respondents (Table 3).

To summarize, wages and salaries of full item non-respondents appear to be underestimated. However, the number of cases that are involved is small, and so statistics that are computed for the full sample and the subset of respondents do not differ much. For self-employment income, instead, full item non-response is very frequent, but we find no evidence of a bias except for in the 10th and 90th percentiles. (Repeating the analysis separately by country and adding regional dummy variables gives similar results.)

5. Conclusions

This paper analyses several issues surrounding income non-response and income imputation, using the ECHP as an illustration.

Comparing final household income for different types of responding households by using UDB 2002 and UDB 2003, we find that relevant improvements have been made to the imputation procedure to take into account unit non-response within responding households. The R^2 for mean and quantile regressions for households with unit non-response increases substantially in UDB 2003 relative to UDB 2002.

Except for the imputation of unit non-responses, there have been no other changes in the imputation procedures between UDB 2002 and UDB 2003, and the results concerning item non-response are similar between the two releases of the data. If we consider the cross-sectional structure of earnings, the imputation procedure for item non-response seems to work well, with the possible exception of a few cases of full item non-response on wages and salaries. For self-employment income, instead, full item non-response is very common but does not appear to lead to significant biases except for the more extreme percentiles.

Acknowledgements

We thank the Associate Editor and two referees for very helpful comments.

Appendix A: Proofs

This appendix collects the proofs of the main results in Section 3.

A.1. Proof 1

(a) To prove that the truncated data estimator is consistent if the assumption of data MCAR (a) holds, note that under model (1)

$$\begin{aligned} E[\psi(X, Y; \theta) D | X] &= E[\psi(X, Y; \theta) | X, D = 1] \Pr(D = 1 | X) \\ &= E[\psi(X, Y; \theta) | X] \Pr(D = 1 | X) = 0. \end{aligned} \tag{10}$$

(b) Following the argument in equation (10), it is straightforward to show that assumption (b) implies equation (3).

(c) Under assumption (c), we have

$$\begin{aligned} E[\psi(X, Y; \theta) D | X] &= E_{X^+}[E[\psi(X, Y; \theta) D | X, X^+]] \\ &= E_{X^+}[E[\psi(X, Y; \theta) | X, X^+, D = 1] \Pr(D = 1 | X, X^+)] \\ &= E_{X^+}[E[\psi(X, Y; \theta) | X] \Pr(D = 1 | X, X^+)] = 0. \end{aligned} \tag{11}$$

A.2. Proof 2

To show that the imputed data estimator is consistent for θ in case (i) ($Z = X$), rewrite the moment restrictions (4) as

$$E[\psi(X, Y; \theta) D | X] + E[\psi(X, Y^*; \theta)(1 - D) | X] = 0, \tag{12}$$

and note that the first term in equation (12) is equivalent to the left-hand side of equation (3) and therefore is equal to 0 under any of the assumptions (a), (b) or (c). The second term is instead equal to

$$E[\psi(X, Y^*; \theta) | X, D = 0] \Pr(D = 0 | X),$$

which is also equal to 0 because of expressions (5) and (1).

A.3. Proof 3

In case (ii), let $Z = (Z_1, Z_2)$, where $Z_1 = X$ and Z_2 are auxiliary variables that are considered in the imputation procedure but excluded from the model of interest. Conditioning and marginalizing the left-hand side of equation (4) with respect to Z_2 gives

$$\begin{aligned} E_{Z_2}[E[\psi(X, Y; \theta) D + \psi(X, Y^*; \theta)(1 - D) | X, Z_2]] &= E_{Z_2}[E[\psi(X, Y; \theta) D | X, Z_2, D = 1] \Pr(D = 1 | X, Z_2) \\ &\quad + E_{Z_2}[\psi(X, Y^*; \theta) | X, Z_2, D = 0] \Pr(D = 0 | X, Z_2)]. \end{aligned} \tag{13}$$

Because of assumptions (d) and (5), $E[\psi(X, Y; \theta)|X, Z_2, D = 1]$ and $E[\psi(X, Y^*; \theta)|X, Z_2, D = 0]$ are both equal to $E[\psi(X, Y; \theta)|Z]$. Hence

$$E_{Z_2}[E[\psi(X, Y; \theta)D + \psi(X, Y^*; \theta)(1 - D)|X, Z_2]] = E_{Z_2}[E[\psi(X, Y; \theta)D|Z_1, Z_2]] = E[\psi(X, Y; \theta)|X] = 0. \quad (14)$$

A similar proof applies when assumptions (5) and (e) hold. (Just condition and marginalize with respect to Z_2 and X^+ rather than with respect to Z_2 only.)

Appendix B: European Community Household Panel imputation of personal income components

Imputation procedures have changed across ECHP releases. In the very first release, imputation was performed by using random hot deck imputation within classes and predictive mean matching. The most recent releases, including UDB 2002 and UDB 2003, adopt a new imputation procedure called imputation and variance estimation (IVE). (This procedure has been carried out by using software that was developed by the Survey Research Center at the Institute for Social Research of the University of Michigan. See Raghunathan *et al.* (1999) for a detailed description.)

When a personal income variable is missing, it is replaced with its most recent observed or imputed lagged value, except for imputation of wages and salaries earnings or when the lagged income variable is also missing, in which case the imputed value is computed by using the IVE procedure.

In the first step of IVE, imputation is applied to variables with a low fraction of missing cases by using the information from variables without missing cases. In the second step, imputation is applied to variables with more severe problems of missingness, conditioning both on variables without missing data and variables imputed in the first step, and so on. The higher 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 that is 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. Imputed values of income variables are forced to lie between the minimum and the maximum values observed for respondents.

For wages and salary earnings and for self-employment income, the imputation model is a log-linear regression with the following explanatory (auxiliary) variables: region of residence, number of workers in the household, age, gender, schooling level, occupation in the current job, status in employment, job status, total number of hours worked per week and main activity and size of the local unit where the person is working. For self-employment income, the marital status of the person and the 'equivalized' household size (using the modified Organisation for Economic Co-operation and Development scale) are also used. The imputation model that is used is a log-linear mean regression model. When some of these variables are missing, as sometimes occurs, they themselves become target variables to be imputed at an earlier stage of the IVE procedure.

In conclusion, the imputation of personal income subcomponents implicitly imposes the dual assumption that $D_{hij,t} \perp\!\!\!\perp Y_{hij,t}|Y_{hij,t-1}$ and $D_{hij,t} \perp\!\!\!\perp Y_{hij,t}|Z_{hi,t}$, where $Y_{hij,q}$ is the jth subcomponent of personal income of individual i in household h in wave q, $D_{hij,t}$ equals 1 when $Y_{hij,t}$ is observed and equals 0 otherwise, and $Z_{hi,t}$ is the set of auxiliary variables that are used in the IVE imputation. The first of these assumptions is necessary because missing values of $Y_{hij,t}$ are replaced systematically by their corresponding lagged values $Y_{hij,t-1}$. The second is instead necessary for the validity of the IVE imputation which uses $Z_{hi,t}$ as explanatory (auxiliary) variables.

If some of the auxiliary variables, say $Z'_{hi,t}$, are missing, they are imputed by using the remaining auxiliary variables, say $Z''_{hi,t}$. This requires the additional assumption that $D_{hij,t} \perp\!\!\!\perp Z'_{hi,t}|Z''_{hi,t}$. This assumption, together with $D_{hij,t} \perp\!\!\!\perp Y_{hij,t}|Z_{hi,t}$, is equivalent to assuming that $D_{hij,t} \perp\!\!\!\perp Y_{hij,t}|Z'_{hi,t}$.

Appendix C: European Community Household Panel imputation for unit non-response within responding households

As for item non-response, the imputation procedures for unit non-response within responding household have changed across releases.

Until UDB 2002, the imputation was carried out by Eurostat for all countries by computing a within-household inflation factor. Construction of this inflation factor starts by computing a 'provisional personal income' for each responding household member, which is equal to the sum of the different types of personal income (reported or imputed), plus the 'assigned' income components (i.e. the value of the income components that were 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 $\bar{Y}_g$ of provisional personal incomes is computed using the cross-sectional weights. (In the ECHP, weights are computed to take account of the sampling design and of non-responding households for which no questionnaire is available. See Eurostat (2002b) for more details.) This weighted average is then assigned to each eligible household member belonging to that group, whether responding or not. Finally, the within-household non-response inflation factor is computed as

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

where $\mathbf{1}\{i \in g\}$ equals 1 if individual i belongs to group g and 0 otherwise, D_{hi} equals 1 if individual i returns the questionnaire and equals 0 otherwise, and Σ_i is the sum over all eligible individuals in household h. To avoid outliers, the within-household non-response factor is set equal to missing if this procedure gives a value that is greater than 5.

In UDB 2003, the imputation procedure changed completely. The new imputation method, which exploits previously neglected information on individual and household income in the current and previous wave, is no longer based on an inflation factor, but on the computation of an 'additional income amount' which is added to Y_h^{I} (the household income after item imputation, as defined in Section 2.4) to allow for unit non-response within the household. (The new imputation method is applied to all countries except Finland, Ireland and the UK, that rely instead on their own imputation methods.) Let Y_h^{m} be the household monthly income reported in the previous wave or, if this information is missing or the composition of the household has changed, the household monthly income reported in the current wave. Then the additional income amount is equal to the difference between Y_h^{m} multiplied by 12 and Y_h^{I}, if this difference is positive, and to 0 otherwise. In other words, the final imputed household income is $Y_h^{\mathrm{F}} = \max(12Y_h^{\mathrm{m}}, Y_h^{\mathrm{I}})$.

In UDB 2002, the imputation for unit non-response within responding households is carried out by implicitly imposing the assumption that $D_h \perp\!\!\!\perp Y_h | Z_h$, where h indexes the household, D_h equals 1 if no household member is a unit non-respondent and 0 otherwise (thus, D_h is the product of the D_{hi} that is defined in Section 2.4 over all members of household h), Y_h is the household income and Z_h is a vector of auxiliary variables. (Specifically, the auxiliary variables that are used are indicators for the age, class and gender of the household members, quintiles of the equivalized monthly household income reported in the current wave and Y_h^{I}.) In UDB 2003, instead, the imputation for unit non-response is carried out by implicitly assuming that $D_h \perp\!\!\!\perp Y_h | Y_h^{\mathrm{m}}$.

If there are no changes in household composition or personal income during the previous year, then it seems sensible to assume that $Y_h = Y_h^{\mathrm{F}} = 12Y_h^{\mathrm{m}}$. Using the information on Y_h^{m} that was observed in the last wave to impute Y_h seems more sensible than using the quintiles of Y_h^{m} that were observed in the current wave and the dummy variables for the gender and age group of the household members. For this reason, the imputation that is adopted in UDB 2003 is probably more appropriate then the imputation that was adopted in UDB 2002.

References

Buchinsky, M. (1998) Recent advances in quantile regression models: a practical guideline for empirical research. *J. Hum. Resour.*, **33**, 88–126.

Cappellari, L. and Jenkins, S. (2004) Modelling low income transitions. *J. Appl. Econometr.*, **19**, 593–610.

Duan, N. (1983) Smearing estimate: a nonparametric retransformation method. *J. Am. Statist. Ass.*, **78**, 605–610.

Eurostat (2002a) *Imputation of Income in the ECHP, PAN 164*. Luxembourg: Eurostat.

Eurostat (2002b) *Construction of Weights in the ECHP, PAN 165*. Luxembourg: Eurostat.

Fay, R. E. (1996) Alternative paradigms for the analysis of imputed survey data. *J. Am. Statist. Ass.*, **91**, 490–498.

Hansen, L. P. (1982) Large sample properties of generalized method of moments estimator. *Econometrica*, **50**, 1029–1054.

Heckman, J. J., Lochner, L. and Todd, P. E. (2003) Fifty years of Mincer earnings regressions. *Working Paper 9732*. National Bureau of Economic Research, Cambridge.
Koenker, R. (2005) *Quantile Regression*. Cambridge: Cambridge University Press.
Koenker, R. and Bassett, Jr, G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
Koenker, R. and Hallock, K. F. (2001) Quantile regression. *J. Econ. Perspect.*, **15**, 143–156.
Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.*, **9**, 538–573.
Mincer, J. (1974) *Schooling, Experience, and Earnings*. New York: Columbia University Press.
Peracchi, F. (2002) The European Community Household Panel: a review. *Empir. Econ.*, **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*. Ann Arbor: Institute for Social Research.
Robins, J. M. and Wang, N. (2000) Inference for imputation estimators. *Biometrika*, **87**, 113–124.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1989) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (1996) Multiple imputation after 18+ years. *J. Am. Statist. Ass.*, **91**, 473–520.
Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
Wooldridge, J. M. (1999) Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica*, **67**, 1385–1406.
Wooldridge, J. M. (2001) Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometr. Theory*, **17**, 451–470.