Poverty Analysis with Unit and Item Nonresponses: Alternative Estimators Compared

Cheti Nicoletti

ISER Working Papers
Number 2003-20

Institute for Social and Economic Research

The Institute for Social and Economic Research (ISER) specialises in the production and analysis of longitudinal data. ISER incorporates the following centres:

- ESRC Research Centre on Micro-social Change. Established in 1989 to identify, explain, model and forecast social change in Britain at the individual and household level, the Centre specialises in research using longitudinal data.

- ESRC UK Longitudinal Centre. This national resource centre was established in October 1999 to promote the use of longitudinal data and to develop a strategy for the future of large-scale longitudinal surveys. It was responsible for the British Household Panel Survey (BHPS) and for the ESRC's interest in the National Child Development Study and the 1970 British Cohort Study

- European Centre for Analysis in the Social Sciences. ECASS is an interdisciplinary research centre which hosts major research programmes and helps researchers from the EU gain access to longitudinal data and cross-national datasets from all over Europe.

The British Household Panel Survey is one of the main instruments for measuring social change in Britain. The BHPS comprises a nationally representative sample of around 5,500 households and over 10,000 individuals who are reinterviewed each year. The questionnaire includes a constant core of items accompanied by a variable component in order to provide for the collection of initial conditions data and to allow for the subsequent inclusion of emerging research and policy concerns.

Among the main projects in ISER's research programme are: the labour market and the division of domestic responsibilities; changes in families and households; modelling households' labour force behaviour; wealth, well-being and socio-economic structure; resource distribution in the household; and modelling techniques and survey methodology.

BHPS data provide the academic community, policymakers and private sector with a unique national resource and allow for comparative research with similar studies in Europe, the United States and Canada.

BHPS data are available from the Data Archive at the University of Essex
http://www.data-archive.ac.uk

Further information about the BHPS and other longitudinal surveys can be obtained by telephoning +44 (0) 1206 873543.

The support of both the Economic and Social Research Council (ESRC) and the University of Essex is gratefully acknowledged. The work reported in this paper is part of the scientific programme of the Institute for Social and Economic Research.

Acknowledgement: Part of this paper is based on work carried out during my visiting at the Institute for Social and Economic Research (ISER), University of Essex. I would like to thank the ISER members for their helpful advice. I am also grateful to seminar and conference participants at the Center for Operations Research and Econometrics (CORE), Univeristé Catholique de Louvain-la-Neuve, at the Innocenzo Gasparini Institute for Economics Research in Milan (IGIER) and at the European Society for Population Economics Conference (ESPE) 2003, New York, for useful comments.

Readers wishing to cite this document are asked to use the following form of words:

> **Familyname, Firstname (August 2003) 'Title of Working Paper', *Working Papers of the Institute for Social and Economic Research*, paper 2003-20. Colchester: University of Essex.**

For an on-line version of this working paper and others in the series, please visit the Institute's website at: http://www.iser.essex.ac.uk/pubs/workpaps/

Institute for Social and Economic Research
University of Essex
Wivenhoe Park
Colchester
Essex
CO4 3SQ UK
Telephone: +44 (0) 1206 872957
Fax: +44 (0) 1206 873151
E-mail: iser@essex.ac.uk
Website: http://www.iser.essex.ac.uk

ABSTRACT

This paper assesses some methods to estimate binary response models in presence of missing data. Particular attention is given to imputation, econometric sample selection correction and propensity score estimation methods. These methods are based on assumptions whose validity can only be verified when the missing data are observed. Using the European Community Household Panel I discuss an informal test to assess the underlying assumptions of different estimation methods for poverty probability. This informal test makes use of Manski bounds for poverty probability, which can be estimated with no or weak assumptions.

NON-TECHNICAL SUMMARY

By estimating poverty probability in Italy in presence of unit and item nonresponses, this paper aims to assess different methods to take account of missing data. The estimation methods applied belong to three types of approaches: imputation, econometric sample selection correction and propensity score approaches. All three types of estimation methods are based on assumptions whose validity can only be verified when the missing data are observed. This is in general impossible by definition of missing data.

Propensity score and imputation methods impose the so called missing at random (MAR) condition, i.e. they impose that the probability of being respondent be independent from unobserved data conditioning on observed data. The econometric selection correction methods relax the MAR condition by substituting it with some assumptions on the relationship between the poverty probability and the nonresponse probability. Both types of assumption are not testable, moreover they are not in general nested. It is therefore impossible to indicate an order of preference between two alternative estimators.

Nevertheless, by applying the estimation procedure proposed by Manski (1989), it is possible to estimate bounds, say Manski bounds, for the poverty probability without imposing any assumption. After the estimation of the Manski bounds, it is possible to check if the estimated probabilities of being poor, using different types of estimation procedures, lie inside the Manski bounds. This constitutes an informal test for the underlying assumptions of alternative estimation procedures.

I apply this type of procedure to compare different estimation methods for the poverty analysis in Italy in 1998 using the European Community Household Panel (ECHP).

1 Introduction

This paper assesses different types of binary model estimation procedures used to take account of missing data. I consider three alternative estimation approaches:

1. the propensity score methods, whose theoretical foundations were introduced by the statisticians Rosenbaum and Rubin (1983) for the evaluation of treatment effects;

2. the econometric sample selection correction methods,[1] adopted by econometricians since the seminal paper of Heckman (1979) and mainly applied to solve endogeneity problems in labour economics;

3. the imputation methods, which are used by survey statisticians to solve the nonresponse problem in sample surveys.

All three types of estimation methods are based on assumptions whose validity can only be verified when the missing data are observed or when adequate data are available to recover the unknown underlying distribution of the missing data. This is for example the case when experimental, simulated or register data are available or when refreshment samples exist besides panel data to solve attrition problems.

Propensity score and imputation methods are based on the assumption that the data are missing at random (MAR), i.e. the probability of being respondent is assumed to be independent from unobserved data conditioning on observed data.[2] On the other side, the econometric selection correction methods relax the MAR condition by allowing the probability of being respondent to depend on both observed and unobserved data.

When the set of observed variables is small and inadequate to describe the probability of being respondent, it seems then reasonable to reject the MAR condition and to prefer the econometric sample selection correction methods. Nevertheless, econometric sample selection correction methods relax the MAR condition at the cost of imposing some other untestable assumptions. In particular, they impose assumptions on the relationship between the error terms in the model of interest and in the model describing the selection. In the case of

[1] Sample selection problems may arise in average treatment effects evaluation using non-experimental data, in impact estimation of an endogenous binary variable on a response variable of interest and in making inference using a sample survey affected by nonresponse, as it is the case in this paper.

[2] The MAR condition is equivalent to the weak unconfoundness, ignorability or conditional independence assumptions, CIA, for the treatment assignment (or program participation).

a parametric approach these assumptions consist in a joint distribution assumption, while in the case of a semiparametric approach they consist in a first moment condition for the dependent variable of interest given the selected sample, which is called separability condition or index restriction.[3] These are not in general nested into the underlying assumptions of the propensity score and the imputation methods. It is therefore impossible to indicate an order of preference among these estimators.

To my knowledge there are only two estimation procedures which are not based on untestable assumptions. The first one is the estimation procedure proposed by Manski (1989) and then extended in some more recent works by Manski (1995), Horowitz and Manski (1998), Manski and Pepper (2000), Vasquez *et al.* (1999 and 2001), and Horowitz *et al.* (2003). It consists in the computation of bounds (henceforth Manski bounds) instead of a point estimation for the specific statistics of interest, generally a conditional mean or quantile. The second one is the estimation method proposed by Hirano *et al.* (2003), that solves the identification problem for panel models, due to attrition, by combining panel data sets with refreshment samples.

I do not consider the procedure of Hirano *et al.* (2003) because a refreshment sample is not available for the panel considered in the empirical application. Furthermore, the focus of the paper is not on attrition but on unit and item nonresponse in a single wave of a panel for individuals belonging to households, for which at least the reference person has been responding. More precisely, the aim of the paper is the poverty analysis in Italy in 1998 using the European Community Household Panel (ECHP) and taking account of unit and item nonresponses on income.

I solve at least partially the identification problem of the probability of being poor by considering the Manski bounds, which estimation does not involve restrictive assumptions. Then, I check if the estimated probabilities of being poor, using different types of estimation procedures, lie inside the Manski bounds. This constitutes an informal test for the underlying assumptions of the different estimation procedures.

An individual is defined to be poor if he/she belongs to a household with a total net income bellow a fixed poverty line. Since the household income variable is given by the sum of subcompenents, each one possibly affected by item or unit nonresponse, in many cases

[3] See Vella (1998) for a survey of parametric and semiparametric econometric approaches to correct for sample selection.

the information on income is not completely absent. I show that this partial information on household income is useful to narrow the width of Manski bounds. Moreover, I introduce some instrumental and monotone instrumental variable assumptions as in Manski and Pepper (2000) to narrow further the Manski bounds. This increases the power of the informal test to verify and compare the underlying assumptions of different estimation methods for the probability of being poor.

The rest of the paper is organized as follows. Section 2 describes different types of point and Manski bounds estimation methods of the poverty probability in presence of missing data. After a brief description of the data used, Section 3 compares and checks the underlying assumptions of different estimation procedures for the analysis of poverty in Italy. Finally, in Section 4, some conclusions are drawn.

2 Poverty analysis in presence of missing data

The most common model used in empirical works to describe the poverty probability as a function of the variables characterizing the persons and their household is the probit model. When the household income, hence the poverty status, is affected by a problem of nonresponse, then the estimation of the probit model disregarding the missing data may be inconsistent. Therefore I consider different estimation procedures of a probit model for the poverty probability, which take account of the possible sample selection due to the missing data. To avoid differences in the estimation procedures linked to differences in the assumptions of the form and specification of the model of interest, I maintain the probit assumption for all the point estimation procedures except for the case of the linear probability model. This assumption can be submitted to a test under the MAR condition and I find out that it is not rejected.

I apply six different types of point estimations:

1. the probit with complete data, i.e. excluding all individuals with a problem of nonresponse on household income,

2. the probit with imputed data, i.e. the estimation of a simple probit model by replacing the unknown poverty dummy which the one computed using an imputed income,

3. the propensity score weighting method, i.e. a probit with weights equal to the inverse

probability of being respondent,

4. a pseudo Cosslett (1991) estimator method, which consists in the estimation of a probit model adding a set of dummy variables indicating different level of propensity to respond,

5. the censored bivariate probit, which models jointly the poverty probability and the response probability allowing the errors to be correlated,

6. the linear probability model with selection (LPM method), which estimates a linear model instead of the probit model for the poverty probability jointly with a probit model for the response probability.

The aim of this paper is then to compare the above estimation methods. There are several papers, which have compared different estimation methods to deal with the missing data problem or with the close problem of the evaluation of treatment effects using non-randomized experiments.

Among papers assessing different estimation methods in presence of missing data there are Verbeek and Njiman (1992) and Jensen *et al.* (2002), which compare different methods to estimate and to test the sample selection caused by attrition in panel data model. Among papers assessing instead the treatment effect evaluation methods there are Heckman *et al.* (1997) and Heckman *et al.* (1999). In all the above papers the assessment of alternative estimation methods is possible because missing data are replaced either by experimental or simulated data.

Without experimental data, simulated data or other data sources to recover the unknown underlying distribution of the missing data, it is not possible to compare and to choose among different types of estimation procedures taking account of the missing data.

Anyway, following the suggestion of Manski (1989), it is possible to informally check the performance of the different estimations by verifying if the estimated values lie outside the Manski bounds computed assuming very weak or no assumptions.

In the next two subsections I describe the six different point estimation methods and the bounds estimation used to compute the probability of being poor.

2.1 Brief description of the estimation methods

Let consider a binary variable, y_i, taking value 1 if the i–th individual is in a poverty status and 0 otherwise. I assume that the binary variable y_i follows a probit model, i.e. I assume that y_i is related to a continuous latent variable y_i^* through the observation rule $y_i = 1\{y_i^* > 0\}$ and this latent random variable y_i^* obeys the regression model

$$y_i^* = x_i\beta + u_i, \tag{1}$$

where x_i are the explanatory variables, β is the parameter vector of interest and u_i are errors independent from x_i identically and independently distributed as a Gaussian with zero mean and unit variance.[4] The latent variable y_i^* is not observed but it is assumed that there exists a proper monotone transformation of it which is equal to the household income variable, say Y. The coefficients β can be estimated by maximizing the log likelihood function for the probit model, say $\ln L$. Given a random sample of n individuals, this implies the following first order condition,

$$\sum_{i=1}^{n} \frac{d}{d\beta} \ln L_i = \sum_{i=1}^{n} \frac{y_i - \Phi(x_i\beta)}{\Phi(x_i\beta)(1 - \Phi(x_i\beta))} \phi(x_i\beta) x_i' = 0, \tag{2}$$

and the following moment condition,

$$E\left(\frac{d}{d\beta} \ln L_i\right) = E\left(\frac{y_i - \Phi(x_i\beta)}{\Phi(x_i\beta)(1 - \Phi(x_i\beta))} \phi(x_i\beta) x_i'\right) = 0. \tag{3}$$

To simplify notation, I will avoid henceforth the subscript i for individuals.

Three definitions of relative poverty are used in this paper: the percentages of people, both children and adults, with a household income below 40%, 50% and 60% of the median household income.[5] The household income is measured as the net equivalized household income (computed by dividing the total net household income by the equivalized household size, OECD modified scale) which, henceforth, I will call briefly household income.

To identify the poverty status of an individual, using the above definitions, we need to observe his/her household income. In other words, any time the household income is missing we cannot observe the dummy variable y indicating the poverty status.

[4] The normalization of the variance is necessary because the coefficients of a binary response model are only identifiable up to scale.

[5] I refer to Smeeding *et al.* (2000) for details on different poverty measures.

Let r be a dummy variable equal to 1 if the household income is observed and 0 if it is partially or not at all observed, that is in the case of a partial or full nonresponse. Then we call *selection process* (mechanism or model) the conditional model describing the probability to observe the household income given a set of observed explanatory variables z and the household income, which may be unobserved, and,

$$f(r \mid Y, z; \gamma) = \Pr(r = 1 \mid Y, z; \gamma)^r \Pr(r = 0 \mid Y, z; \gamma)^{1-r}. \tag{4}$$

The set of variables z may include some or all the explanatory variables, x, involved in the model of interest. These variables may be personal and household characteristics as well as data collection characteristics, which are important in explaining the probability of nonresponse. Without any loss of generality let $x = (x^c, x^y)$, $z = (x^c, x^r)$ and $w = (x^c, x^y, x^r)$. The explanatory variables w are assumed to be observed for all individuals.

If the probability that $r = 1$ does not depend on any of the observed or unobserved characteristics, i.e.

$$\Pr(r = 1 \mid Y, x^c, x^y, x^r; \gamma) = \Pr(r = 1), \tag{5}$$

then we say that the data are *missing completely at random* (MCAR). In that case we can ignore the missing data problem and estimate the probit model using the subsample of individuals with observations on both y and x, say the *truncated sample*.

If the probability that $r = 1$ depends on observed characteristics but does not depend on unobserved ones, then we say that the data are *missing at random* (MAR). In other words, the data are MAR when the probability to observe Y depends on the set of observed variables, x^c, x^y, x^r, but does not depend on Y, say

$$r \perp\!\!\!\perp Y \mid x^c, x^y, x^r. \tag{6}$$

MAR alone does not ensure the consistency of estimators based on the truncated sample. This is because of the potential *selection on observables*, as defined in Heckman and Hotz (1989) and in Heckman (1990). Nevertheless, the selection on observables vanishes in the following two cases:

- if we estimate the model of interest by controlling for all relevant explanatory variables in the selection mechanism,
- if the exclusion restriction of variables x^r from the model of interest (x^r are variables relevant for the selection model) is satisfied.

Henceforth I will call the above exclusion restriction condition *instrumental variables (IV) exclusion restriction*. If we do not consider x^r among the explanatory variables of the model of interest and the IV exclusion restriction is not satisfied, then we must adopt an estimation method correcting for the selection on the observables.

Finally if the probability that $r_i = 1$ depends on both observed and unobserved characteristics, then the sample selection is informative and a proper estimation method correcting for it must be adopted. Using the terminology of Heckman and Hotz (1989), we say that there is selection on unobservables.

In the following we describe the different methods used in the empirical application, some ignoring the selection process, some correcting only for the selection on observables and some other correcting for the selection on unobservables too.

The first method considered is the estimation of a probit model for the poverty probability disregarding the individuals with missing data, say briefly the ignoring method. This method does not consider any type of selection and it is consistent when the MAR condition,

$$r \perp\!\!\!\perp Y \mid x^c, x^y, x^r, \tag{7}$$

and the IV exclusion restriction,

$$u \perp\!\!\!\perp x^r, \tag{8}$$

are satisfied.

The second solution adopted to solve this missing data problem is to replace the missing household incomes by imputed values[6] in the estimation of a probit model for the poverty probability. This estimation procedure, say imputation method, takes account of selection on observables but does not consider the selection on unobservables. Indicating with y^I the imputed poverty dummy the consistency of the imputation method is ensured if either the MAR assumption is satisfied and $E(y^I \mid w) = E(y \mid w)$, or if $(r \perp\!\!\!\perp Y \mid x)$ and $E(y^I \mid x) = E(y \mid x)$. The conditions $(r \perp\!\!\!\perp Y \mid x^c, x^y, x^r)$ and $(r \perp\!\!\!\perp Y \mid x^c, x^y)$ are not nested, but are equivalent when $(Y \perp\!\!\!\perp x^r \mid x^c, x^y)$ or $(r \perp\!\!\!\perp x^r \mid x^c, x^y)$.[7]

[6] I use the imputed data available in the user database of the European Community Household Panel (ECHP-UDB), which is the dataset used for the empirical analysis. A brief description of the imputation method adopted in the ECHP-UDB is given in the appendix.

[7] Proofs of the statements reported in this section are available on request to the author. Note that if the imputation procedure replaces the missing value of the household income, Y, with a value Y^I such that $E(Y^I \mid w) = E(Y \mid w)$, then this does not ensure that the imputed poverty is such that $E(y^I \mid w) = E(y \mid w)$. This is because y is not a linear function of Y.

The third method used to estimate the probit in presence of missing data is the propensity score weighting method (say briefly weighting method). This method consists in the estimation of a probit model with weights given by the inverse probability to observe Y given the set of relevant observed variables,

$$\Pr(r = 1 \,|\, w) = p(z; \gamma), \tag{9}$$

say the inverse propensity score.[8] This type of estimation is usually called propensity score weighting estimation and has its roots in the inverse probability weighted estimator proposed by Horvitz and Thompson (1952). Horvitz and Thompson (1952) use it to compute population means taking account of the variation of sampling and response probabilities across units of the population. This idea has been recently reconsidered by Robins and Rotnitzky (1995), Robins *et al.* (1995) and Abowd *et al.* (2001) for the estimation of conditional means in the presence of missing data and by Rosembaum and Rubin (1983), Imbens (2000) and Hirano *et al.* (2000) for the evaluation of treatment effects. The inverse propensity score weighting method is consistent if the MAR condition holds.

The forth estimation method for the poverty probability model is a modified version of the Cosslett (1991) estimator. The Cosslett estimator is a semiparametric sample selection correction method, which imposes the separability condition. The separability condition is the condition allowing to write the regression equation (1) for the truncated sample as

$$y^* = x\beta + E(u \,|\, r = 1, p(z; \gamma)) + \nu, \tag{10}$$

where ν is a residual error term with mean zero and

$$E(u \,|\, r = 1, p(z; \gamma)) = g(p(z; \gamma)). \tag{11}$$

To avoid assumptions on the form of the function $g((p(z; \gamma))$ I substitute it with a set of dummy variables indicating the subsets of a partition of the support of the propensity score.[9]

[8] Since the propensity score is replaced by its estimated value, we should take account of that in the computation of the variance of the estimator. A possible way to consistently compute the estimator variance is by estimating jointly the selection model and the model of interest as in Abowd, Crépon and Kramarz (2001). Nevertheless, in the application we focus attention only on the bias of the estimator, while we do not consider the estimation of its variance.

[9] The partition of the sample is performed by dividing the [0,1] support of the propensity in equally spaced subintervals and by controlling that the balancing score properties is satisfied (see Rosenbaum and Rubin 1983 for a definition of this property).

Note that this estimation procedure has a strong relation with the propensity score stratification methods, which also control for the selection bias by stratifying the sample in s disjoint sub-samples associated with s disjoint subintervals of the $[0,1]$ support of the propensity score.

It is then possible to estimate consistently the parameter β by regressing the latent variable y^* on the set of explanatory variables and the above dummy variables. Unfortunately in our case y^* is unobserved. Furthermore, since the residual error term ν is not anymore a normal error, the consistency of the Cosslett estimator applied to a probit model is not ensured. Nevertheless, Horowitz (1993) studied the effects of a distributional misspecification in quantal response model. His results seem to suggest that, when the true density is unimodal and homoskedastic, the distributional misspecification errors are small as long as the incorrect density distribution is also unimodal and homoskedastic. For this reason we think that allowing for the variance of ν to depend on the set of the above dummy variables may reduce the bias caused by the ditributional misspecification.

The fifth and the sixth estimation procedures belongs to the parametric sample selection correction approach, which takes account of the selection on unobservable by allowing the error terms in the model of interest and in the selection process to be correlated. These two methods impose a joint distribution for the errors, while the Cosslett estimator, which belongs instead to the semiparametric sample selection correction approach, imposes only the separability condition. For a recent review of the parametric and semiparametric econometric selection approaches I refer to Vella (1998).

The fifth method is given by the maximum likelihood estimation of a censored bivariate probit model, i.e. both the model of interest and the selection model are latent index model with errors distributed as a bivariate normal with means 0, unit variances and covariance different from 0. The sixth method is given instead by the maximum likelihood estimation of a censored regression model, say a generalized Tobit model. In other words the sixth method replace the probit model for the poverty probability with a linear probability model. Both the fifth and the sixth methods assume some IV exclusion restriction, more precisely they assume that x^r are irrelevant for the poverty model.

In the empirical application I assume a probit model for the selection process in the propensity score weighing method, in the pseudo Cosslett method and in the censored bivariate probit model. I verify the normality assumption for the error, which is not rejected.

To verify the normality of the errors, I use the score test for normality of the error in an ordered probit presented in Machin and Stewart (1990), which is a modification of the score test derived by Chesher and Irish (1987) for a grouped dependent variable. Obviously, I modify the test derived by Chesher and Irish (1987) to consider a binary dependent variable instead of an ordered categorical.

This test is also used to verify the normality assumption of the errors in the poverty model under the MAR condition. Again the normality of the errors in the poverty model is not rejected.

Notice that when the MAR and the IV exclusion restriction hold, the probit model ignoring the missing data, the pseudo Cosslett estimator and the propensity score estimation are consistent even when the selection process (propensity score) is estimated omitting the IV, x^r. This is because the MAR and the IV exclusion restriction imply that $u \perp\!\!\!\perp r \mid w$ and $u \perp\!\!\!\perp x^r \mid x$ so that the probit model ignoring the missing data is consistent as well as the pseudo Cosslett estimation, which considers some unnecessary dummy variables to approximate the zero value function $E(u \mid r = 1, p(z;\gamma)) = g(p(z;\gamma)) = 0$. The maximum likelihood estimation of the probit model for the poverty probability with weights given by the inverse propensity score $p(x^c;\theta)$ instead of $p(x^c, x^r;\gamma)$ is also consistent. This is because the moment condition implied by the first order condition for the maximum likelihood maximization continues to hold as proved below. Let us consider the moment condition for the score function with weights given by the inverse propensity score $p(x^c;\theta)$:

$$E\left(\frac{d}{d\beta}\ln L\frac{r}{p(x^c;\theta)} \mid x\right) = E\left(\frac{y-\Phi(x\beta)}{\Phi(x\beta)(1-\Phi(x\beta))}\phi(x\beta)x'\frac{r}{p(x^c;\theta)} \mid x\right), \tag{12}$$

and let us condition and marginalize the above expression with respect to x^r,

$$E_{x^r}\left[E\left(\frac{y-\Phi(x\beta)}{\Phi(x\beta)(1-\Phi(x\beta))}\phi(x\beta)x' \mid w, r=1\right)\frac{Pr(r=1 \mid w)}{p(x^c;\theta)}\right]. \tag{13}$$

Since $y \perp\!\!\!\perp r \mid w$ and $y \perp\!\!\!\perp x^r \mid x$ we can write again the above moment condition as

$$\begin{aligned} &E\left(\frac{y-\Phi(x\beta)}{\Phi(x\beta)(1-\Phi(x\beta))}\phi(x\beta)x' \mid w, r=1\right)E_{x^r}\left[\frac{Pr(r=1 \mid w)}{p(x^c;\theta)}\right] = \\ &= E\left(\frac{y-\Phi(x\beta)}{\Phi(x\beta)(1-\Phi(x\beta))}\phi(x\beta)x' \mid x\right)E_{x^r}\left[\frac{Pr(r=1 \mid w)}{p(x^c;\theta)}\right] = 0, \end{aligned} \tag{14}$$

so that the consistency of propensity score estimation method continue to hold.

Omitting the IV from the sample selection in the censored bivariate probit or in the linear probability model with selection can instead cause some problem of "empirical" identification, so that these estimators can give biased results even if the MAR and the IV exclusion restriction hold.

2.2 Bounds estimation for the poverty probability

Following the approach used in Manski (1995), Horowitz and Manski (1998), Manski and Pepper (2000), Vasquez *et al.* (1999, 2001), I compute bounds for the poverty probability without imposing any assumption on the missing data or by imposing some weak assumptions to reduce the width of the bounds. Furthermore, I try to narrow the bounds by using available information on the partial reported income for partial nonresponding households.

The idea behind the computation of bounds for a probability, such as the probability of being poor, is simple and has been introduced by Manski (1989). Let r be a dummy taking value 1 if an individual belongs to a household with complete response on income (i.e. a responding household whose total income is fully reported) and 0 otherwise, let Y be his/her household income, and let c be the poverty line. In general the probability of being poor, $\Pr\{Y < c\}$, cannot be identified when Y is known only for a subsample of the individuals. By using the law of total probability, it is possible to decompose the probability of poverty as

$$\Pr\{Y < c\} = \Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\} + \Pr\{Y < c \,|\, r = 0\} \Pr\{r = 0\}, \qquad (15)$$

and we can identify 3 of the 4 elements in the right hand side of the above equation. The unknown element is $\Pr\{Y < c \,|\, r = 0\}$, which takes values between 0 and 1. We can therefore compute an upper and a lower bound (henceforth UB and LB) for the probability of poverty by substituting to the unknown element the maximum and the minimum values in its support, i.e.

$$\begin{aligned} UB &= \Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\} + \Pr\{r = 0\}, \\ LB &= \Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\}. \end{aligned} \qquad (16)$$

These bounds are usually called the "worst case" bounds.

Since the household income is given by the sum of the personal incomes of each household member, which in turn are given by the sum of different personal income subcomponents, it often occurs that some of the household income subcompoenents are missing and other

are observed, so that the household income can be observed only partially. Most of the households that are not responding give a partial information on their income, i.e., we know a reported household income which consists in a lower threshold for the household income. Let Y^r be the partially reported value of the household income (which is 0 in the case of a full item nonresponse), then, by using again the law of total probability, we can decompose the unknown probability as follow:

$$\begin{aligned}\Pr\{Y < c \,|\, r = 0\} = \Pr\{Y < c \,|\, Y^r < c, r = 0\} \Pr\{Y^r < c \,|\, r = 0\} + \\ + \Pr\{Y < c \,|\, Y^r \geq c, r = 0\} \Pr\{Y^r \geq c \,|\, r = 0\}.\end{aligned} \tag{17}$$

Since Y is always greater or equal to Y^r, it follows that $\Pr\{Y < c \,|\, Y^r \geq c, r = 0\} = 0$ and so the second addend on the right-hand side of the above equation cancels out. Because we know the reported household income for the nonresponding individuals, we can estimate $\Pr\{Y^r < c \,|\, r = 0\}$. The exact value of the probability $\Pr\{Y < c \,|\, Y^r < c, r = 0\}$ is instead unknown, but it lies between 0 and 1. This allows to compute the following new upper bound, which I call reported income upper bound,

$$UB_r = \Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\} + \Pr\{Y^r < c \,|\, r = 0\} \Pr\{r = 0\}. \tag{18}$$

The information on reported income does not affect instead the lower bound, which remains unchanged with respect to the worst case bound, LB. The use of the partial reported income allows to narrow the width of the bounds from $\Pr\{r = 0\}$ to $\Pr\{Y^r < c \,|\, r = 0\} \Pr\{r = 0\}$. The data on partially reported incomes have not been used before to narrow the Manski bounds and I show in the empirical section that the use of this additional information may be very useful to solve at least partially the identification problem of the probability of being poor.

Furthermore, I impose different types of weak assumptions that can help narrowing further the Manski bounds. In particular I introduce some instrumental variable and some monotone instrumental variable assumptions (see Manski 1995 and Manski and Pepper 2000 for more details).

I use a dummy indicating the use of the same interviewer for the same household across waves as an instrumental variable (IV). This means that I assume that the poverty probability, conditioning on a set of covariates, is independent from the dummy variable indicating the use of the same interviewer, which is assumed instead to be relevant for the nonresponse probability.

Moreover, I assume that the poverty probability, conditioning on a set of covariates, is monotonically increasing with the household size and monotonically decreasing with the number of workers. In other words, using the terminology of Manski (1995) and Manski and Pepper (2000), the size of the household and the number of workers are assumed to be monotone instrumental variables (MIV).

Let Z be the IV and X be the set of conditioning variables,[10] then $\Pr\{Y < c \,|\, X, Z\} = \Pr\{Y < c \,|\, X\}$ and the bounds for $\Pr\{Y < c \,|\, X, Z\}$ are also bounds for $\Pr\{Y < c \,|\, X\}$ so that

$$\begin{aligned} LB_{IV} &= \sup_z \Pr\{Y < c \,|\, X, Z, r = 1\} \Pr\{r = 1 \,|\, X, Z)\} \\ &\leq \Pr\{Y < c \,|\, x\} \\ &\leq \inf_Z \Pr(Y < c \,|\, X, Z, r = 1\} \Pr\{r = 1 \,|\, X, Z\} + \Pr\{r = 0 \,|\, X, Z\} \\ &= UB_{IV}, \end{aligned} \tag{19}$$

where sup and inf indicate the supremum and the infimum. I call these bounds IV lower bound, LB_{IV}, and IV upper bound UB_{IV}.

If Z is instead a MIV, then we know that $\Pr\{Y < c \,|\, X, Z = z_1\} > \Pr\{Y < c \,|\, X, Z = z_2\}$ whenever $z_1 > z_2$ (when for example the MIV is the household size) or whenever $z_1 \leq z_2$ (when for example the MIV is the number of workers). Taking as example the case of the number of members in the household, then the bounds for $\Pr\{Y < c \,|\, X, Z\}$, say MIV bounds, are given by

$$\begin{aligned} LB_{MIV} &= \sup_{v > z_1} \Pr\{Y < c \,|\, X, Z = z_1, r = 1\} \Pr\{r = 1 \,|\, X, Z = z_1\} \\ &\leq \Pr\{Y < c \,|\, X, Z = v\} \\ &\leq \inf_{v < z_2} \Pr\{Y < c \,|\, X, Z = z_2, r = 1\} \Pr\{r = 1 \,|\, X, Z = z_2\} + \Pr\{r = 0 \,|\, X, Z = z_2\} \\ &= UB_{MIV}. \end{aligned} \tag{20}$$

I call these bounds MIV lower bound, LB_{MIV}, and MIV upper bound UB_{MIV}.

The covariates X are characteristics of the household, of the reference person in the household and of the data collection process. More precisely I consider: (i) a dummy indicating age of the reference is between 40 and 65 years, (ii) a dummy for the low level of education (less than second stage of secondary education), (iii) a dummy indicating the use of the same interviewer across waves, (iv) the number of workers in the household, (v) the size of the household.

[10] I use the uppercase for the variables x and z to distinguish them from the variables used in the models for probability of being poor and the probability of being respondent in the last section.

I estimate nonparametrically the IV bounds and the two MIV bounds conditioning to the set of all the above variables, X. The bounds for the marginal poverty probability are then computed by integrating out the conditioning variables using the law of total probability.

Before describing the results of the bounds estimates, there is a consideration worth noting. When using the imputed income values, the estimated poverty probability lies always inside of both the worse case bounds and of the reported bounds. As stressed by Horowitz and Manski (1998), "estimates using imputations take the observed data as given and specify logically possible values for the missing data. Thence imputation always yields a logically possible value of the conditional expectation of interest". In particular, this is true for donor imputation methods, but it might be false for model imputation methods. In the ECHP the imputed values are constrained to be between the minimum and the maximum values observed for the responding individuals, so that the imputed income takes only logically possible values, under the assumption that the household income has a common support for respondents and nonrespondents.

In our case we are interested in a dummy indicating the poverty status and the imputed values for the missing household income, say Y^I, are obviously such that $0 < \Pr\{Y^I < c \,|\, r = 0\} < 1$. Thence the imputed poverty probability, that is, the probability computed replacing missing incomes with their imputed values,

$$\Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\} + \Pr\{Y^I < c \,|\, r = 0\} \Pr\{r = 0\}, \tag{21}$$

always lies between the lower and the upper worst case bounds.

The bounds computed using the reported income narrow, but the imputed poverty probability remains inside the bounds. This is because the imputed values are always greater or equal to the reported values, thence $0 < \Pr\{Y^I < c \,|\, r = 0\} < \Pr\{Y^r < c \,|\, r = 0\}$.

If the imputed values are used instead to replace the missing values in the estimation of a probit model for poverty, then the predicted probabilities may lie outside of the Manski bounds. The estimation of a probit model using the imputed values may lead to unconsistent estimation of the parameters, which are used to predict the poverty probability. This is because the conditions for the consistency of the estimator may fail.

3 Poverty analysis: empirical results

In this section I carry out the poverty analysis for Italy in 1998 using the ECHP UDB 2002 (the User Data Base of the European Community Household Panel Survey release 2002), which is an anonymized and user-friendly version of the ECHP data. In Section 3.1 I describe briefly the ECHP data and the types of nonresponse which may affect the income. Furthermore, I give the definitions of household income and poverty used in the empirical application. Section 3.2 shows the results of different estimation procedures of the probability of being poor. Finally, Section 3.3 presents the results of the Manski bounds estimation and of the informal check of the underlying assumptions of the different estimation methods applied.

3.1 Brief description of the data

The ECHP is a standardized multi-purpose annual longitudinal survey carried out for the 15 European countries belonging to the European Union (EU). It is centrally designed and coordinated by the Statistical Office of the European Communities (Eurostat). A more detailed description of the ECHP can be found in Peracchi (2002) and Eurostat (2001a).

The target population of the ECHP consists of all individuals living in private households within the EU. In its first (1994) wave, the ECHP covered about 60,000 households and 130,000 individuals aged 16+ in 12 countries of the EU (Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain and the UK). Austria, Finland and Sweden began to participate later. I focus attention only on Italy and on the last wave available, which refers to the 1998.

For the empirical analysis I define different measures of relative poverty by using the total net household income. In the ECHP the total net household income is obtained by summing over the different types of income and over the individuals belonging to a same household and it is measured in annual amounts in the year before the survey, net of taxes and expressed in national units and current prices. In the application, to allow comparability across different types of households, the household income is measured as the net equivalized household income at constant 1995 prices (I use the equivalized size, OECD modified scale).

Different types of nonresponse may affect the household income, these may be classified in:

1. household unit nonresponse, when no household members give back the questionnaire, neither the personal questionnaire nor the household questionnaire;

2. personal unit nonresponse for some of the members of the households, when some persons in the household give back the questionnaire, but some other are unit nonresponding;

3. personal item nonresponse, when one or more members give back the questionnaire but they do not answer to all questions on the specific income components.

The household income nonresponses may be classified in fully and partial nonresponses. If at least one income sub-component is known for at least one member of the household, then there is a partial nonresponse. Full nonresponse occurs instead if a household is unit nonresponding or if all members of the household do not answer to any of the income questions despite being possibly unit respondents. For households affected by income nonresponse it is possible to observe partially the household income, say the *reported income*, which will be 0 in the case of a full nonresponse and higher than 0 in the case of a partial nonresponse.

In the empirical application I use the reported income to solve at least partially the identification of the poverty probability using Manski bounds, see Section 2.2. I focus attention on the poverty probability in 1998 for people belonging to households, for which at least the reference person returned the personal questionnaire and the household questionnaire. In other words I exclude the households defined at the point (1). In this way it is possible to use the information on the households and on their reference persons to explain the probability of a nonresponse (partial or full) on the household income. The size of the sample used is of 16746 individuals, of which 3288 have a missing household income (20%). The households unit nonresponding are excluded from the analysis.

Three definitions of poverty are used: the percentages of people with income below 40%, 50% and 60% of the median income (see Smeeding *et al.* 2000 for definitions of poverty in a cross-national context). I compute the median income and the poverty probability separately by country using all members (both children and adults) of responding households in the ECHP. The median income is computed using the imputed value and the weights provided in the ECHP-UDB to take account of personal item and unit nonresponses and of household unit nonresponses. Obviously the estimation of the median income may be affected by inconsistencies in the imputation and weighting procedures adopted in the ECHP. This may

have an impact on the estimation of the poverty line, but it should not have any consequence on the comparison of the poverty estimation procedures, for which I use the same poverty line.

3.2 Estimation of a poverty model

In this section I present the results of the six following different types of estimations, already described in Section 3.2:

1. the probit with imputed data (imputation),
2. the propensity score weighting method (weighting),
3. the censored bivariate probit (joint),
4. the probit with complete data (ignoring),
5. the pseudo Cosslett method (Cosslett),
6. the linear probability model with selection (LPM).

For the specification of the poverty and selection models I follow Cappellari and Jenkins (2002). To choose the variables affecting the response probability I follow also some other recent papers assessing the item and unit nonresponse behaviour, in particular D'Alessio and Faiella (2002), Fitzgerald *et al.* (1996), Hill and Willis (2001), Lepkowski and Couper (2002), Lillard and Panis (1998), Riphahan and Serfling (2002) and Schrápler (2002). Using the ECHP-UDB it is impossible to distinguish between household members item nonresponding and fully responding, I consider therefore a dummy for the response defined at household level. This dummy takes value 1 for all individuals belonging to a household with a fully response on household income and 0 for all individuals belonging to a household with a partial or fully nonresponse on household income.

All estimations are obtained by allowing the error terms to be correlated for individuals belonging to the same household.

I consider the three relative measures of poverty defined in the last section, and I use as explanatory variables, x, in the poverty probit model the following ones:

- dummies for the age of the individuals and of the reference person in the household (two dummies, one for age between 40 and 65 and one for age higher than 65),

- indicators for the highest level of completed education of the reference person (two dummies, one for college and one for a level of education lower than secondary one),
- the size of the household measured by the number of members,
- two dummies for the presence of 1, and 2 or more children,
- a dummy for the sex gender of the reference person,
- a dummy for a reference person without a spouse,
- a dummy for the home tenure,
- indicators of the labor status of the reference person (inactive, unemployed, self-employed),
- the number of workers in the household.

In addition to the above variables, to explain the probability to respond, I use the following ones, x^r:

- the mode of interview (one dummy to distinguish face to face interviews with respect to telephone and self-administered interviews),
- a dummy to indicate if the individual belongs to the original sample drawn in the first wave of the panel,
- the number of visits of the interviewer to the household,
- a dummy indicating the use of the same interviewer for the same household across waves.

The above variables are linked to the collection process and are likely to affect the probability to respond but should not affect the probability of being poor. In other words these are the IV which are assumed to be irrelevant for the poverty probability when considering the censored bivariate probit, the linear probability model with selection, or the probit with complete data.

Submitting to a test the probit specification for the selection model, I find that the normality assumption for the error is not rejected.

All estimations results are similar in terms of sign and significance of the coefficients. The most important variables in explaining poverty are:

- the age dummies (there is a positive relationship between the probability of being poor and the presence of young people whether reference people or other members of the households);
- the number of workers, which is negatively related to poverty;
- the household size, the dummies for the presence of children, the dummy indicating a level of education lower than the secondary one, the dummy for a reference person without a spouse and the indicator of self-employed status, which are all positively related to poverty.

The additional variables used as explanatories for the probability to respond are adequate IV, indeed, they are not relevant in explaining the poverty probability, at least under the MAR condition. Nevertheless, in the selection model only the interview mode indicators and the dummy for the use of the same interviewer across waves are significantly different from 0 when using a significance level equal to 0.05.

The assumption of a zero correlation between errors in the censored bivariate probit is not rejected so that the MAR assumption is not rejected, at least under the joint distributional assumption. Furthermore a Hausman type test to verify the equality of the inverse propensity score weighted and the unweighted probit (the probit with complete data) estimators does not reject the null hypothesis. Under the assumption that the data are MAR, the above Hausman type test allows to conclude that the exclusion restriction of the IV from the poverty equation is not rejected.

In conclusion it seems that it is possible to make inference for the poverty model disregarding the missing data, if we are willing to accept the joint distributional assumption. If we are instead willing to accept the MAR assumption it seems that the IV exclusion restriction and the probit specification for the poverty probability are not rejected. Unfortunately it is not possible to verify the probit specification assumption and the IV exclusion restriction without imposing the MAR condition, and viceversa. Nevertheless an informal check to verify the validity of the underlying assumptions of different types of estimation can be conducted by checking if the estimated probabilities of being poor lie inside the Manski bounds. The results of this checking procedure are reported in the following section.

3.3 Comparison of the estimation procedures

Table 1 shows the worst case bounds estimates (LB and UB) and their confidence intervals (CI lower for the lower bound and CI upper for the upper bound), the upper bound estimated using the reported income (UB_r) and the corresponding upper confidence interval band (CI upper). The bounds are computed for three alternative definitions of poverty line, namely 40%, 50% and 60% of median income. The confidence intervals are computed by bootstrap (1000 samples with replacement are drawn from the original data) and by taking the 5th percentile and the 95th percentile of the bootstrap distribution for the corresponding lower and upper confidence bands.

Using reported income does help in narrowing the bounds. Indeed, the reported upper bound is always much lower than the worst case upper bound. The length of the interval between the upper and the lower bound narrows down from about 20 to about 7 percentage points.

Because the width of the confidence intervals is much narrower than the width of the bounds, finding weak assumptions to narrow the bounds is much more important than increasing the sample size to reduce sampling variability. I introduce then the IV and the MIV assumptions. More precisely, I use as IV the dummy indicating the use of the same interviewer across waves, and two monotone instrumental variables, says MIV1 and MIV2, the number of worker and the household size. The bounds are computed conditioning to the set of variables X, defined in the last section, and then integrating out the conditioning variables using the law of total probability. The new bounds are shown in Table 2. Their width shrinks slightly with respect the reported bounds.

Table 3 reports the predicted poverty probabilities using different types of estimation. All methods predict poverty probabilities higher than the one computed using the imputed income variables, except obviously the estimation of the probit model using imputed values. It seems therefore that the imputation procedure leads to a slight underestimation of the poverty probability. Moreover, the linear probability model seems to produce a slight overestimation of the poverty probability.

In Table 4 I report the percentages of inconsistencies, i.e. the percentages of cases in which the conditional predicted poverty probabilities lie outside the Manski bounds confidence intervals. The percentage of inconsistencies for each type of estimation and each type of Manski bounds are computed as follow:

- the sample of individuals is divided in cells, $s = 1, ..., S$, using the variables X defined in Section 2.2, each one containing a subsample of individuals of size N_s,
- for each cell s the upper (lower) bound for the probability of being poor is estimated by the naïf nonparametric estimation, by using the formulas reported in Section 2.2 and replacing the theoretical probabilities by the empirical frequencies,
- for each cell s the predicted probability of being poor is computed using the estimated coefficients (of the specific type of estimation considered) and setting the explanatory variables to the corresponding values for the cell,
- the lower (upper) confidence band for each lower (upper) Manski bound is estimated by bootstrap considering 1000 replications and taking the 5th (95th) percentile of the corresponding distributions,
- for each cell s I define a dummy d_s^c taking value 1 if the estimated probability of being poor lies between the lower and the upper confidence band and 0 otherwise,
- finally the percentage of inconsistencies for each estimator is given by

$$\frac{(\sum_s d_s^c N_s) \cdot 100}{\sum_s N_s}. \tag{22}$$

I consider acceptable the estimators with a percentage of inconsistencies lower or equal than 10%. All the estimators seem to be acceptable except the linear probability model, which seems to be inadequate to describe a binary model. The censored bivariate probit model has some problems when using a poverty line defined as the 40% of the median income for the bounds computed using the IV and the MIV. This may be due to an identification problem, which may affect this estimator when the IV used in the selection models are not very significant. The imputation method presents also some problems in the case of a poverty line defined as 60% of the median income and using the IV. Since the dummy indicating the poverty y is not a linear function of the household income Y, $E(Y^I \mid w) = E(Y \mid w)$ does not imply $E(y^I \mid w) = E(y \mid w)$. For this reason the poverty probability estimated using imputed values may be biased even under MAR and consistency of the procedure adopted by ECHP to impute household income. Nevertheless, I would conclude that all estimators perform well except the linear probability model.

Then I focus the attention on the consequences of omitting relevant variables (the number of workers and the dummies indicating the labor status of the reference person) from both the model of interest and the selection process, only from the model of interest and only from selection model. I study also the consequences of the omission of the IV (mode of interview, dummy for the use of the same interviewer across waves, dummy for the individuals belonging to the original sample, number of visits) from the selection model.

When omitting some relevant variables from both equations the estimators perform badly, i.e. the percentage of inconsistencies has a dramatic increase, see Table 5. The only exceptions occur for the censored bivariate model, which seems to perform well for a poverty line defined as 40% and 50% of the median income and when disregarding the information on the reported income in computing Manski bounds. It seems that the bias caused by the omission of some explanatory variables from both equations may be partially corrected by allowing the error terms to be correlated. I find indeed that the correlation between the errors is about -0.80 and significantly different from 0 at 1% level for all three definitions of poverty line.

The same dramatic increase in the inconsistencies occurs when omitting relevant variables from the main equation of interest. In this case the censored bivariate model performs badly too.

When instead relevant variables are omitted only from the selection equation, see Table 6, the propensity score weighting and the Cosslett estimators produce predicted values still consistent with the Manski bounds. This is indeed a reasonable result when the MAR condition is valid and the IV are not relevant for the main equation, see Section 3.2.

The probit models estimated using the imputed values and disregarding the units with missing data are obviously not affected by changes in the estimation of the selection process.

Finally the censored bivariate probit and the linear probability model have several inconsistencies when omitting relevant variables from the selection process.

I obtain the same result when eliminating the IV (mode of interview, dummy for the use of the same interviewer across waves, dummy for the individuals belonging to the original sample, number of visits) from the selection equation. The omission of important variables or IV for the selection model seems to cause an identification problem for the parametric sample selection correction methods (the censored bivariate probit and the linear probability model with selection), while do not affect the other estimators. As a result of the changes in the

specification of the selection model, the correlation between the errors becomes significantly different from 0. It seems therefore that, under the MAR condition, the parametric sample selection correction methods may be seriously affected by the possible misspecification of the selection model, while the other methods seem to be more robust.

4 Conclusion

The empirical results suggest that, even when the percentage of nonresponses on household income is high, it is possible to narrow significantly the width of the Manski bounds for the poverty probability by using information on partially reported income. In the application the household income is missing in about 20% of cases, this would imply a gap between the upper and the lower Manski bound of 20 percentage points, which is not very informative. Using instead the partially reported income it is possible to narrow the width of the Manski bounds from 20 to 7 percentage points.

Furthermore, it seems that, using the partial information on household income, the Manski bounds are enough informative to run an informal check of the underlying assumptions of different types of estimators. In particular, it seems possible to detect the inconsistency of an estimator by checking if its predicted poverty probabilities lie inside the Manski bounds. Obviously the check is an informal test for which the power is not known. Nevertheless, in the empirical application this informal check seems to work well in detecting cases in which the estimators are inconsistent. When excluding important explanatory variables from the poverty model, all the estimation methods perform very badly in terms of number of predicted values lying inside the Manski bounds. Furthermore, the results show that there may be a serious identification problem when trying to estimate a censored bivariate probit or a censored regression model without proper instrumental variables. The absence of IV does not seem instead to affect the weighting propensity score estimation and the pseudo Colsslett's estimation, which perform well even when some explanatory variables are excluded from the selection model. This is an expected result when the data are MAR and the IV exclusion restriction for the poverty model holds.

In conclusion, the Manski bounds can provide a powerful informal test to verify the untestable underlying assumptions of the imputation, econometric sample selection correction and propensity score methods. In particular in my empirical application the Manski bounds seem to support the assumptions of MAR and IV exclusion restriction.

References

Abowd J., Crépon B., Kramarz F. (2001), "Moment estimation with attrition", *Journal of the American Statistical Association*, 96, 456, 1223-1231.

Cappellari L, Jenkins S. (2002), "Modelling low income dynamics", *Working Papers of the Institute for Social and Economic Research*, paper 2002-08, Colchester: University of Essex.

Chesher A., Irish M. (1987), "Residual analysis in the grouped and censored normal linear model", *Journal of Econometrics*, 34, 33-61.

Cosslett S.R. (1991), "Semiparametric estimation of a regression model with sample selectivity", in W.A. Barnett, J. Powell, G.E. Tauchen (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press, New York.

D'Alessio G., Faiella I. (2003), Non-response behaviour in the Bank of Italy's survey of household income and wealth, *Banca d'Italia, Temi di discussione*, 462.

Eurostat (2001a), *ECHP UDB Manual, Waves 1 to 5.*

Eurostat (2001b), "Imputation of income in the ECHP", PAN 164/01.

Fitzgerald J., Gottschalk P., Moffitt R. (1996), An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics, *The Journal of Human Resources*, 33, 251–299.

Heckman J. (1979), "Sample selection as a specification error", *Econometrica*, 47, 1, 153–161.

Heckman J. (1990), "Varieties of selection bias", *The American Economic Reivew*, 80, 2, 313–318.

Heckman J., Hotz V.J. (1989), "Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training", *Journal of the American Statistical Association*, 84, 408, 862–874.

Heckman J., Ichimura H., Smith J., Todd P. (1998), " Characterizing selection bias using experimental data", *Econometrica*, 66, 5, 1017–1098.

Heckman J.J., LaLonde R.J., Smith J.A. (2000), " The economics and econometrics of active labor mearket programs", in O. Ashenfelter and D. Card, (eds.), *Handbook of Labor Economics*, 3, North Holland, Amsterdam.

Hill D., Willis R. (2001), "Reducing panel attrition: a search for effective policy instruments", *The Journal of Human Resources*, 36, 3, 416-438.

Hirano K., Imbens G.W., Ridder G. (2000), "Efficient estimation of average treatment effects using the estimated propensity score", NBER working papers, 251.

Hirano K., Imbens G.W., Ridder G., Rubin D.R. (2003), "Combining panels with attrition and refreshment samples", *Econometrica*, 69, 6, 1645-1659.

Horowitz J.L. (1993), " Semiparametric and nonparametric estimation of quantal response models", in Maddala C.R. and Vinod H.D. eds., *Handbook of Statistics*, 11, 45–72.

Horowitz J.L., 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.

Horowitz J.L., Manski C.F., Ponomareva M., Stoye J. (2003), "Computation of bounds on population parameters when data are incomplete", forthcoming in *Reliable Computing*.

Horvitz D., Thompson D. (1952), "A generalization of sampling without replacement from a finite population", *Journal of the American Statistical Association*, 47, 260, 663–685.

Imbens G.W. (2000), "The role of the propensity score in estimating dose-response functions", *Biometrika*, 87, 3, 706–710.

Jensen P., Rosholm M., Verner M. (2002), "A comparison of different estimators for panel data sample selection models", Department of Economics, University of Aarhus, Working paper, 2002-1.

Lepkowski J.M., Couper M.P. (2002), "Nonresponse in the second wave of longitudinal household surveys", in R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds.), *Survey Nonresponse*, Wiley, New York.

Lillard L.A., Panis C.W.A. (1998), Panel attrition from the panel study of income dynamics, *The Journal of Human Resources*, 33, 439–547.

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

Machin S.J., Stewart M.B. (1990), "Union and financial performance of British private sector establishment", *Journal of Applied Econometrics*, 5, 4, 327-350.

Manski C.F. (1989), "Anatomy of the selection bias", *The Journal of Human Resources*, 24, 3, 343–360.

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

Manski C.F., and Pepper J.V. (2000), "Monotone instrumental variables: with an application to return to schooling", *Econometrica*, 68, 4, 997–1010.

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

Raghunathan T.E., Solenberger P.W., 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.

Riphahn R.T., Serfling O. (2002), "Item non-response on income and wealth questions", *IZA Discussion Paper*, 573.

Robins J., Rotnitzky A. (1995), "Semiparametric efficiency in multivariate regression models with missing data", *Journal of the American Statistical Association*, 90, 429, 122–129.

Robins J., Rotnitzky A., Zhao L. (1995), "Analysis of semiparametric regression models for repeated outcomes in presence of missing data", *Journal of the American Statistical Association*, 90, 106-121.

Rosenbaum P.R., Rubin D.B. (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70, 1, 41–55.

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

Schafer J.L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Schräpler J.-P. (2002), "Respondent behavior in panel studies. A case study for income-nonresponse by means of the German Socio-Economic Panel (GSOEP)", *German Institute for Economic Research Discussion Paper*, 299, Berlin.

Smeeding T., Rainwater L., and Burtless G. (2000), "United Stated poverty in a cross-national context", in S.H. Danziger and R.H. Haveman (eds), *Understanding Poverty*, Russel Sage Foundation and Harvard University Press, New York and Cambridge, MA.

Vazquez R., Melenberg B., and van Soest A. (1999), "Bounds on quantiles in the presence of full and item nonresponse", CentER Discussion Paper, 1999–38, Tilburg University.

Vazquez R., Melenberg B., and van Soest A. (2001), "Nonparametric bounds in the presence of item nonrespone, unfolding brackets, and anchoring", CentER Discussion Paper, 2001–67, Tilburg University.

Vella F. (1998), "Estimating models with sample selection bias: a survey", *The Journal of Human Resources*, 33, 1, 127-169.

Verbeek M., Nijman T. (1992), "Testing for selectivity bias in panel data models", *International Economic Review*, 33, 3, 681-704.

Table 1: Worst case and reported income bounds.

Bounds	Poverty 40%	Poverty 50%	Poverty 60%
Imputed poverty	7.6	12.3	18.8
Poverty for respondents	7.8	13.0	19.9
Imputed poverty for nonrespondents	5.8	8.2	12.2
CI lower LB	6.1	10.3	16.0
Lower bound (LB)	6.5	10.7	16.5
Upper bound reported (UBr)	11.4	16.6	23.6
CI upper Ubr	11.9	17.1	24.3
Upper bound (UB)	26.1	30.4	36.1
CI upper UB	26.8	31.1	36.8

Table 2: Poverty probabilities bounds using IV and MIV.

Bounds	Poverty 40%	Poverty 50%	Poverty 60%
Imputed poverty	7.6	12.3	18.8
Poverty for respondents	7.8	13.0	19.9
Imputed poverty for nonrespondents	5.8	8.2	12.2
lbIV	7.6	12.3	18.1
ubIV	24.5	28.8	34.6
ubrIV	10.6	15.5	22.1
lbMIV n. workers	6.5	10.8	16.5
ubMIV n. workers	24.0	29.3	35.5
ubrMIV n. workers	11.0	16.3	23.5
lbMIV household size	7.0	11.2	16.9
ubMIV household size	25.2	29.8	35.6
ubhrMIV household size	11.1	16.3	23.4

Table 3: Predicted poverty probabilities using different estimation methods.

Bounds	Poverty 40%	Poverty 50%	Poverty 60%
Imputed poverty	7.6	12.3	18.8
Poverty for respondents	7.8	13.0	19.9
Imputed poverty for nonrespondents	5.8	8.2	12.2
Probit with imputed data	7.6	12.3	18.8
Propensity score wighting	8.2	13.4	20.4
Censored bivariate probit	8.8	13.6	19.8
Probit with complete data	8.2	13.4	20.3
Pseudo Cosslett	8.3	13.4	20.3
Linear probability model with selection	8.7	14.1	20.9

Table 4: Percentages of inconsistencies with respect to the bounds.

	Poverty line 40% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB, UB)$	1.2	1.2	1.2	1.2	1.2	9.6
$out(LB_{IV}, UB_{IV})$	3.2	3.2	3.2	3.2	3.2	9.6
$out(LB_{MIV1}, UB_{MIV1})$	1.2	1.2	1.2	1.2	1.2	9.6
$out(LB_{MIV2}, UB_{MIV2})$	1.2	1.2	1.2	1.2	1.2	9.6
$out(LBr, UBr)$	1.2	1.6	11.9	1.6	8.3	17.1
$out(LBr_{IV}, UBr_{IV})$	3.2	3.2	13.5	3.2	13.5	20.6
$out(LBr_{MIV1}, UBr_{MIV1})$	1.2	2.7	15.4	1.6	8.3	20.6
$out(LBr_{MIV2}, UBr_{MIV2})$	1.2	1.6	11.9	1.6	8.3	17.1
	Poverty line 50% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB, UB)$	2.0	0.9	2.0	2.0	0.9	8.0
$out(LB_{IV}, UB_{IV})$	3.6	2.5	3.6	3.6	2.5	9.6
$out(LB_{MIV1}, UB_{MIV1})$	2.0	0.9	2.0	2.0	0.9	8.0
$out(LB_{MIV2}, UB_{MIV2})$	2.0	0.9	2.0	2.0	0.9	8.0
$out(LBr, UBr)$	2.4	1.3	2.4	2.4	8.4	16.2
$out(LBr_{IV}, UBr_{IV})$	4.3	3.2	4.3	4.3	10.3	19.6
$out(LBr_{MIV1}, UBr_{MIV1})$	2.4	2.8	3.9	3.9	8.4	17.7
$out(LBr_{MIV2}, UBr_{MIV2})$	2.4	1.3	2.4	2.4	8.4	16.8
	Poverty line 60% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB, UB)$	5.9	1.1	1.7	1.1	0.6	2.7
$out(LB_{IV}, UB_{IV})$	15.2	2.7	4.0	3.4	0.6	4.3
$out(LB_{MIV1}, UB_{MIV1})$	5.9	1.1	1.7	1.1	0.6	2.7
$out(LB_{MIV2}, UB_{MIV2})$	5.9	1.1	1.7	1.1	0.6	8.0
$out(LBr, UBr)$	6.2	1.8	2.5	1.8	8.1	6.2
$out(LBr_{IV}, UBr_{IV})$	15.9	4.6	5.0	4.4	9.3	13.3
$out(LBr_{MIV1}, UBr_{MIV1})$	6.3	1.8	2.5	1.8	8.1	6.2
$out(LBr_{MIV2}, UBr_{MIV2})$	6.3	1.8	2.5	1.8	8.1	11.5

Table 5: Percentages of inconsistencies with respect to the bounds when relevant variables are omitted from the poverty and the selection models.

	Poverty line 40% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB, UB)$	17.5	17.5	1.2	17.5	17.5	7.6
$out(LB_{IV}, UB_{IV})$	18.3	18.3	3.2	18.3	18.3	8.4
$out(LB_{MIV1}, UB_{MIV1})$	17.5	17.5	6.0	17.5	17.5	7.6
$out(LB_{MIV2}, UB_{MIV2})$	18.3	18.3	2.7	18.3	18.3	8.4
$out(LBr, UBr)$	29.8	40.9	81.6	40.9	40.9	31.0
$out(LBr_{IV}, UBr_{IV})$	36.7	53.3	87.4	53.3	44.1	43.4
$out(LBr_{MIV1}, UBr_{MIV1})$	34.1	45.3	85.3	45.3	45.3	35.4
$out(LBr_{MIV2}, UBr_{MIV2})$	36.2	47.3	88.2	47.3	47.3	37.4
	Poverty line 50% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB, UB)$	20.8	20.8	2.5	20.8	20.8	20.8
$out(LB_{IV}, UB_{IV})$	32.8	32.8	4.0	32.8	32.8	32.8
$out(LB_{MIV1}, UB_{MIV1})$	20.8	20.8	5.8	20.8	20.8	20.8
$out(LB_{MIV2}, UB_{MIV2})$	21.5	21.5	2.9	21.5	21.5	21.5
$out(LBr, UBr)$	44.2	49.6	75.7	49.6	44.2	51.1
$out(LBr_{IV}, UBr_{IV})$	67.8	67.8	81.7	67.8	67.8	67.8
$out(LBr_{MIV1}, UBr_{MIV1})$	44.2	49.6	79.4	49.6	46.6	53.5
$out(LBr_{MIV2}, UBr_{MIV2})$	50.6	56.0	83.2	56.0	50.6	60.1
	Poverty line 60% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB, UB)$	34.9	23.4	20.4	23.4	23.4	21.9
$out(LB_{IV}, UB_{IV})$	44.0	37.0	26.8	37.0	39.7	35.4
$out(LB_{MIV1}, UB_{MIV1})$	34.9	23.4	22.8	23.4	23.4	21.9
$out(LB_{MIV2}, UB_{MIV2})$	35.7	24.2	27.5	24.2	23.8	22.2
$out(LBr, UBr)$	60.2	57.2	57.2	63.4	58.1	61.9
$out(LBr_{IV}, UBr_{IV})$	70.2	75.4	69.0	77.3	76.7	79.2
$out(LBr_{MIV1}, UBr_{MIV1})$	60.2	57.2	64.9	58.1	58.1	61.9
$out(LBr_{MIV2}, UBr_{MIV2})$	71.6	68.5	69.2	71.4	73.8	76.5

Table 6: Percentages of inconsistencies with respect to the bounds when relevant variables are omitted from the selection model.

	Poverty line 40% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB,UB)$	1.2	1.2	0.0	1.2	1.2	8.0
$out(LB_{IV},UB_{IV})$	3.2	3.2	4.1	3.2	3.2	9.6
$out(LB_{MIV1},UB_{MIV1})$	1.2	1.2	0.0	1.2	1.2	8.0
$out(LB_{MIV2},UB_{MIV2})$	1.2	1.2	1.8	1.2	1.2	8.0
$out(LBr,UBr)$	1.2	1.6	64.5	1.6	1.6	18.7
$out(LBr_{IV},UBr_{IV})$	3.2	6.4	80.8	3.2	3.2	20.6
$out(LBr_{MIV1},UBr_{MIV1})$	1.2	2.7	70.4	1.6	1.6	22.2
$out(LBr_{MIV2},UBr_{MIV2})$	1.2	5.4	73.4	1.6	4.3	25.2
	Poverty line 50% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB,UB)$	2.0	2.0	0.0	2.0	0.9	8.0
$out(LB_{IV},UB_{IV})$	3.6	3.6	0.0	3.6	2.5	9.6
$out(LB_{MIV1},UB_{MIV1})$	2.0	2.0	0.0	2.0	0.9	8.0
$out(LB_{MIV2},UB_{MIV2})$	2.0	2.0	2.7	2.0	2.7	8.0
$out(LBr,UBr)$	2.4	2.4	67.2	2.4	1.3	23.3
$out(LBr_{IV},UBr_{IV})$	4.3	4.3	85.9	4.3	3.2	32.7
$out(LBr_{MIV1},UBr_{MIV1})$	2.4	2.4	71.1	3.9	1.3	24.8
$out(LBr_{MIV2},UBr_{MIV2})$	2.4	6.5	75.9	2.4	5.4	28.9
	Poverty line 60% median income					
	Imputation	Weighting	Joint	Ignoring	Cosslett	LPM
$out(LB,UB)$	5.9	1.7	0.0	1.1	5.0	2.7
$out(LB_{IV},UB_{IV})$	15.3	4.0	0.0	3.4	6.6	4.3
$out(LB_{MIV1},UB_{MIV1})$	5.9	1.7	0.0	1.1	5.0	2.7
$out(LB_{MIV2},UB_{MIV2})$	5.9	4.5	7.3	1.1	7.7	10.8
$out(LBr,UBr)$	6.3	2.5	76.2	1.8	5.4	21.3
$out(LBr_{IV},UBr_{IV})$	15.9	5.0	91.6	4.4	7.3	35.7
$out(LBr_{MIV1},UBr_{MIV1})$	6.3	2.5	76.2	1.8	5.4	22.8
$out(LBr_{MIV2},UBr_{MIV2})$	6.3	19.3	83.3	1.8	19.0	37.9

A Imputation of the income variables in the ECHP

To solve the problem of item nonresponse to income questions, Eurostat applies an imputation procedure at the individual level to compute the missing personal income components.

The way in which household income is computed depends on the presence of unit nonresponse within the household. For households where all eligible members returned their questionnaire, household income is simply obtained by adding up the reported or imputed values of their personal income components. For households with unit nonresponse, namely those where some household members did not return the questionnaire, household income is obtained in three steps. In the first step, the personal incomes of each item nonresponding member are imputed as described below. In the second step, "imputed household income" Y_h^I is computed as the sum of reported and imputed incomes of responding household members, that is,

$$Y_h^I = \sum_i D_{hi}[R_{hi}Y_{hi} + (1 - R_{hi})\widehat{Y}_{hi}], \tag{23}$$

where $\sum_i$ denotes summation over all eligible members of household h, D_{hi} equals 1 if individual i returns the questionnaire and 0 otherwise, R_{hi} equals 1 if individual i answers all questions on personal income and 0 otherwise, and Y_{hi} and $\widehat{Y}_{hi}$ are respectively reported and imputed personal income. In the third step, "final household income" Y_h^F is computed by inflating the imputed household income Y_h^I through a "within-household nonresponse inflation factor" $f_h > 1$. All components reported at the personal level are multiplied by this factor.

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 cross-sectional weights to take account of the unit nonresponses. This weighted average is then assigned to each eligible household member belonging to that group, whether responding or not.

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

$$f_h = \frac{\sum_g \overline{Y}_g \sum_i 1\{i \in g\}}{\sum_g \overline{Y}_g \sum_i 1\{i \in g\} D_{hi}}, \tag{24}$$

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.

Eurostat computes the income, $\widehat{Y}_{hi}$, for item nonrespondent individuals using an imputation procedure called IVE (Imputation and Variance Estimation), [11] which may be viewed as a variant of the EM algorithm (see e.g. Little and Rubin 1987, Rubin 1989 and Schafer 1997 for more detail on the EM procedures), 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. The imputation procedure proceeds by steps. In the first step, imputation is applied to variables with a low fraction of missing cases and uses 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. In the initial stage, the auxiliary variables are sex, age, employment characteristics (socio-professional category, employment sector, size of the firm, type of job, hours worked per week) and education level. Even these variables are sometimes missing, and so they become target variables to be imputed at a previous step of the IVE procedure. For the imputation of a specific target variable past information may also be used. In particular, the value observed for the target variable in the previous wave is used as an auxiliary variable for the imputation of its current value, but not for the imputation of other variables. If the value of the target variable in last wave is not observed but imputed, it is not used.

[11] The imputation has been carried out using the Imputation and Variance Estimation (IVE) software, developed by the Survey Research Center at the Institute for Social Research of the University of Michigan (for a description see Eurostat 2001b and Raghunathan, Solenberger and Hoewyk 1999).

The IVE procedure allows defining a range for the variable to be imputed. In the ECHP this range is equal to the observed range for responding people, that is imputed value must lie between the minimum and the maximum values observed for the responding persons.