Correcting for sample selection bias in poverty analysis: Alternative estimators compared

Cheti Nicoletti
ISER, University of Essex

Preliminary draft

January 2002

Abstract

This paper assesses three different types of regression estimation procedures used to take account of sample selection problems, in particular the missing data problem in sample surveys. These are propensity score estimation, imputation and classical econometric selection models procedures. All three types of estimation methods are based on assumptions whose validity can only be verified when the missing (counterfactual) data are observed.

Nevertheless, by computing bounds instead of a point estimate, it is possible to avoid untestable assumptions and to carry out an informal check of the underlying assumptions of the above estimators, as suggested in Manski (1989). The check procedure involves two steps. The first step consists of the computation of bounds, say Manski bounds, for a specific feature of interest in a regression model, for example the conditional mean or a conditional quantile, with very weak or no assumptions on the missing data mechanism. The second step consists of checking whether the estimates of interest, using alternative estimation methods, lie inside the Manski bounds.

This checking procedure is applied to the estimation of the poverty probability in Italy using the European Community Household Panel Survey. Poverty is defined by using net household income, which is affected by nonresponse in more than 20% of the cases. Such a high nonresponse rate implies that Manski bounds on the probability of being poor tend to be wide. In many cases, however, the information on income is not completely absent because income may be reported partially, i.e. it is known that total net household income is above a known threshold. I use this information on partial reported income and some weak assumptions to narrow Manski bounds. I then check whether the conditional poverty probabilities, estimated by using different methods, contradict the Manski bounds.

Contents

1 Introduction

This paper assesses different types of regression estimation procedures used to take account of sample selection problems. These problems typically 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. I focus my attention on this last case and I consider three alternative estimation approaches to take account of the nonresponse problem. These are:

1. the propensity score methods, which theoretical fundaments were introduced by the statisticians Rosenbaum and Rubin (1984) for the evaluation of treatment effects;

2. the classical econometric selection models, adopted by econometricians since the milestone 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 (for example using experimental data) or when adequate data are available to replace the missing data (for example in panel data analysis when a refreshment sample is available, see Hirano, Imbens, Ridder and Rubin, 1998).

Propensity score and imputation methods are based on the assumption that the data are missing at random (MAR), i.e. the probability of the selection of the sample is independent from unobserved data conditioning on the observed data.[1] On the other side, the econometric selection models relax the MAR condition by allowing the sample selection rule to depend on both observed and unobserved data.

When the set of observed variables is small and inadequate to describe the selection mechanism, it seems then reasonable to reject the MAR condition and to prefer the selection models procedures to other types of estimations. Nevertheless, selection models estimation methods relaxing the MAR condition impose some other untestable assumptions. These are not in general nested into the underlying assumptions of the weighting propensity score and the imputation methods. It is therefore impossible to indicate an order of preference among these estimators.

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

Much effort has been devoted to clarifying the relationships between assumptions imposed by different estimation procedures correcting for the selection bias, in particular for the treatment effects evaluation, see Heckman (1990), Angrist, Imbens and Rubin (1996), Imbens and Angrist (1994), Angrist (1997) and Vytlacil (2002). Though the strong connections between the estimation procedures applied to treatment effect evaluations and regression models with missing data, there are a few differences. For this reason this paper presents an analysis of the assumptions imposed by regression estimation procedures in the special case of the missing data problem. This analysis is useful in understanding whether it is possible to define an order of preference or an equivalence result between estimators; but it cannot help in verifying the validity of untestable assumptions.

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, Melenberg and van Soest (1999 and 2001), and Horowitz and Manski (2002). 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, Imbens, Ridder and Rubin (1999), which 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 (1999) because in the empirical application I conduct a cross-sectional analysis focusing attention on nonresponse in a single wave of a panel rather than on attrition. Furthermore, refreshment samples are not available in the panel used in the application. I instead consider the Manski bounds and use them as a tool to assess the quality of the three above different estimation procedures. More precisely, I check if the three estimations are in agreement with the Manski bounds computed by avoiding untestable assumptions or by imposing very weak assumptions. Moreover, I show that when the dependent variable of interest is given by the sum of subcompenents, each one possibly affected by item nonresponse, the partial information of the aggregate variable is very useful to shrink the Manski bounds.

I apply the estimation procedures to poverty analysis in Italy using the European Community Household Panel survey (ECHP). The empirical analysis focuses on the estimation of a probit model for a dummy indicating the poverty status, but the theoretical part considers also continuous regression models.

The rest of the paper is organised as follows. Section 2 describes the three above estimation procedures and the differences and the relationships of their underlying assumptions. Section 3 applies different estimation procedures to the analysis of poverty in Italy with missing data; while Section 4 assess these estimations by checking whether the predicted poverty probabilities lie inside the Manski bounds. Finally, some conclusions for future developments are drawn Section 5.

2 Estimation methods with missing data

In this section I describe different estimation methods to take account of missing data and their underlying assumptions. In Section 2.1 I define a general parametric model of interest and a generalized method of moments (GMM) estimation procedure to estimate its parameters, which is consistent in the absence of missing data. I prove that the consistency continue to hold in the presence of missing data when the data are missing completely at random (MCAR), i.e. when the probability of sample selection does not depend neither on the observed nor on the unobserved variables.

In Section 2.2, I relax the MCAR assumption and I show that the consistency of the GMM estimator ignoring missing data continues to hold under two sets of different assumptions. Both sets of assumptions require that data are missing at random (MAR), i.e. the probability of sample selection does not depend on the unobserved variables given the observed variables. However, the MAR alone does not ensure the consistency of the GMM estimator. This is due to the problem of selection on observable variables.

Next, in Section 2.3, I consider three estimation procedures assuming MAR and correcting for the possible selection on observables. These methods are the weighted GMM estimator using as weights the inverse probability of selection (i.e. the inverse propensity score), the GMM estimator replacing the missing data with values given by an imputation procedure, and the GMM estimator corrected by using a control function. I prove that the underlying assumptions of the weighted GMM estimator are included in the sets of assumptions imposed by the other two methods.

If the MAR condition is not satisfied, then we have also a problem of selection on unobservables. In that case the above estimators are not consistent. The estimators, which have been proposed to take account of the selection on unobservables, are the following:

1. the ML estimator of a joint parametric model, which consider jointly a model of interest and a selection model, which is the parametric selection models approach,

2. the semiparametric selection models approach, the stratified and matching propensity score methods to estimate the slope coefficients of a linear regression model,

3. the semiparametric selection models approach, the stratified and matching propensity score methods to estimate both the slope and the intercept coefficients of a linear regression model.

Then, in Sectionunobservablesm I consider the above estimators.

Finally, in Section2.5 I describe the main findings concerning the relationships between assumptions imposed by different estimation procedures in the presence of sample selection. I explain which are the costs of relaxing the assumptions imposed by some estimators and which are the testable and the untestable assumptions. Furthermore, I emphasize when it is possible to establish an order of preference between two estimators because the underlying assumptions of one are included in the underlying assumptions of the other. A summary of these findings are presented in Table7.

This section follows the structure of the Table 7 . The first top block of the Table 7 considers the results about the GMM estimation ignoring the missing data, see Section 2.1 and 2.2. Then, the Table 7 reports synthetically the underlying assumptions and the costs of the estimators taking account of selection on observables, which are presented in the Subsection 2.3. Finally, it presents the assumptions and the costs of the estimators taking account of selection on unobservables, which are described in more details in the Subsection 2.4.

2.1 General statement and defintions

Let us assume we are interested in making inference on a conditional model for a variable y belonging to the sample space $\mathcal{Y}$, given a set of variables x belonging to the space $\mathcal{X}$, say

$$\{\mathcal{Y}, f(y \mid x; \theta), \theta \in \Theta\},$$

where $f(y \mid x; \theta)$ is a family of conditional probability distributions indexed by the parameter θ, and Θ is the parameter space. Furthermore, assume that the true data generating process is the joint model

$$\{\mathcal{Y} \times \mathcal{W} \times \mathcal{R}, f(y, w, r; \varphi), \varphi \in \Phi\},$$

where r is a binary variable, $\mathcal{R}$ is its sample space and w is a vector of random variables, which includes the variables y and x in the model of interest.

Assume that a random sample of size N of realisations for (w, r) is observed, while y is observed only if $r = 1$, say in $n < N$ cases. Furthermore, assume that the sample space for the missing

(counterfactual) values is equal to the space for the observed values, $\mathcal{Y}$, i.e. the probability that $r = 1$ is always strictly lower than 1 and strictly higher then 0 for all possible values of y (common support assumption).

In the following I call the subsample of units for which y is observed the truncated sample. I call the process generating the dummy r the selection model (process or mechanism), which can be in a general form defined as

$$\{\mathcal{R}, f(r \mid y, z\gamma), \gamma \in \Gamma\}.$$

Furthermore, I assume that the sets of variables x and z have a subset of common variables, say x^c, and a subset of distinct variables, say x^y and x^r, i.e.

$$x = (x^y, x^c), z = (x^r, x^c), w = (x^y, x^c, x^r).$$

The selection mechanism describes the probability that a unit is selected in the sample. In the case of the missing data problem the selection mechanism (missing data mechanism) describes the probability that a unit is respondent. In the case instead of the evaluation of treatments or programs effects, the selection mechanism describes the probability that a unit is treated or is participating to a program.

To avoid any misunderstanding, I state here briefly the conditions, which allow making correct inference on the conditional model of interest ignoring the selection mechanism. I refer to Rubin (1976) for a more formal presentation of these conditions and to Nicoletti (2002) for an extension to the dynamic panel models.

I focus attention on the generalised method of moment estimation (GMM) procedure, which is enough general to include most of the estimation methods used in econometrics such as the maximum likelihood estimation (ML), the least squares estimation (LS) and the instrumental variable estimation (IV).[2]

Assume that the set of moment conditions used to estimate the parameter θ of the model of interest is:

$$E\left[\psi(y, x^y, x^c; \theta) \mid x^y, x^c\right] = 0. \tag{1}$$

where $\psi(.)$, say the moment condition function, is a vector function with size greater or equal to the size of the parameter vector θ and 0 is a vector of zeros. $\psi(.)$ is equal to the likelihood score function in the case of a ML estimator, it is equal to product of the error and the set of explanatory

[2] See the seminal paper of Hansen (1982) for an analysis of the properties of the GMM estimator and the Mátyás (1999) book for a recent survey on the GMM estimator.

variables for a regression model in the case of the LS estimator and it is equal to the product of the instrumental variables and the error for a regression model in the case of the IV estimator.

Assume that the GMM estimator, say $\hat{\theta}_{GMM}$, which uses the moment conditions in (1), is consistent when y is observed for the full random sample. In the following I give the conditions under which the consistency remains valid in the presence of missing data.

Let us say that data are *missing completely at random* (MCAR), if r is independent of both observed and unobserved variables,

$$r \perp\!\!\!\perp (y, x^y, x^c, x^r).$$

If the data are MCAR, then it is possible to make correct inference on the joint model $f(y, x^y, x^c, x^r)$ and on any other model, which is an admissible reduction of this joint model.[3]

Under MCAR it is therefore possible to ignore the selection mechanism and to estimate the conditional model of interest applying the above GMM estimator, $\hat{\theta}_{GMM}$, to the truncated sample. The moment conditions for the truncated sample are still equal to a vector of zeros and indeed it is possible to write

$$E\left[\psi(y, x^y, x^c; \theta) r \,|\, x^y, x^c\right] = E\left[\psi(y, x^y, x^c; \theta) \,|\, x^y, x^c, r = 1\right] Pr(r = 1 \,|\, x^y, x^c) = 0, \qquad (2)$$

where the last equality is a consequence of the MCAR condition, which implies that y and r are independent conditioning to (x^y, x^c).

Before beginning to describe the different estimation relaxing the MCAR assumption, I give some definitions, which will be useful in the following.

Data are *missing at random* (MAR) if

$$y \perp\!\!\!\perp r \,|\, (x^y, x^c, x^r).$$

Data are missing completely at random given (x^y, x^c), say briefly MCAR $\,|\,(x^y, x^c)$ or *conditional MCAR*, if

$$(y, x^r) \perp\!\!\!\perp r \,|\, (x^y, x^c).$$

We call the condition $(y \perp\!\!\!\perp x^r \,|\, x^y, x^c)$ *instrumental variables exclusion restrictions* (IV exclusion restrictions). The instrumental variables have not to do with the IV estimator mentioned above. The instrumental variables are the set of variables x^r, which enter in the selection process but are excluded (irrelevant) in the model of interest.

[3] An admissible model reduction requires some adequate statistical cuts (initial cuts and sequential cuts in the case of repeated observations). I refer to Engle, Hendry and Richard (1983) for further details on it.

Requiring the MCAR is equivalent to requiring the MAR plus two additional assumptions: $r \perp\!\!\!\perp x^r \,|\, (x^y, x^c)$ and $r \perp\!\!\!\perp (x^y, x^c)$. The conditional MCAR is instead equivalent to the MAR plus $r \perp\!\!\!\perp x^r \,|\, (x^y, x^c)$. Therefore, the MCAR condition is equivalent to the conditional MCAR condition plus $r \perp\!\!\!\perp x^r \,|\, (x^y, x^c)$.

2.2 Estimation methods relaxing MCAR assumptions

Under the MCAR condition it is possible to make correct inference on the joint model $f(y, x^y, x^c, x^r)$ and on any other model, which is an admissible reduction of this joint model, using the truncated sample. If we are interested in estimating the parameter θ of the conditional model

$$\{\mathcal{Y}, f(y \,|\, x; \theta), \theta \in \Theta\},$$

then the MCAR condition is a sufficient but not a necessary condition for a consistent inference on θ based on the generalized method of moment estimation, $\hat{\theta}_{GMM}$. It can be indeed substituted with the weaker assumption of conditional MCAR.

Furthermore, if the IV exclusion restriction is valid, $(y \perp\!\!\!\perp x^r \,|\, x^y, x^c)$, then the MCAR $|\, (x^y, x^c)$ condition can be replaced by the MAR one, i.e. $(r \perp\!\!\!\perp y \,|\, x^y, x^c, x^r)$.

Under the MAR condition we can indeed write

$$E\left[\psi(y, x^y, x^c; \theta) r \,|\, x^y, x^c\right] = E_{x^r}\left[E\left[\psi(y, x^y, x^c; \theta) \,|\, x^y, x^c, x^r, r = 1\right] Pr(r = 1 \,|\, x^y, x^c, x^r)\right], \qquad (3)$$

Where ψ is the moment function introduced in the last section. The above moment condition is equal to 0 if either $(y \perp\!\!\!\perp x^r \,|\, x^y, x^c)$ or $(y \perp\!\!\!\perp x^r \,|\, x^y, x^c)$ is true.

The MAR condition alone is not sufficient for a correct inference neglecting the selection process. This is because

$$E\left[\psi(y, x^y, x^c; \theta) r \,|\, x^y, x^c, x^r\right] \neq E\left[\psi(y, x^y, x^c; \theta) \,|\, x^y, x^c\right].$$

Using the terminology of Heckman and Hotz (1989), this is a problem of selection on observables.

Note that, when all variables relevant to explain the selection process are either included or irrelevant for the conditional model of interest, the problem of selection on observables does not occur. In that situation the MAR condition is a sufficient and necessary condition to make correct inference on θ by applying the GMM estimator with the truncated sample.

In conclusion, the GMM estimator using the truncated sample is consistent under 3 sets of different assumptions:

1. the MCAR condition, $r \perp\!\!\!\perp (y, x^y, x^c, x^r)$, which is equivalent to the MAR condition plus two additional conditions $(r \perp\!\!\!\perp x^r \mid x^y, x^c)$ and $(r \perp\!\!\!\perp x^y, x^c)$;

2. the conditional MCAR, i.e. $(r \perp\!\!\!\perp y, x^r \mid x^y, x^c)$, which is equivalent to the MAR condition plus the condition $(r \perp\!\!\!\perp x^r \mid x^y, x^c)$;

3. the MAR condition, i.e. $(r \perp\!\!\!\perp y \mid x^y, x^c, x^r)$, plus the IV exclusion restriction $(y \perp\!\!\!\perp x^r \mid x^y, x^c)$.

Note that the MAR assumption is not testable, but some of the above conditions are testable. When the MAR assumption holds, it is possible to verify the exclusion restrictions. Furthermore, since (r, x^y, x^c, x^r) are supposed to be always observed, it is possible also to test the conditions $(r \perp\!\!\!\perp x^r \mid x^y, x^c)$ and $(r \perp\!\!\!\perp x^y, x^c)$.

The MAR alone implies a problem of selection on observables, which is usually solved by applying one of the following estimators:

1. the inverse probability weighted GMM estimator or the propensity score weighting method,

2. the GMM estimator corrected by considering a control function.

3. the GMM estimator using imputed data.

2.3 Selection on observable variables

As already said in last section, the MAR condition alone does not ensure a consistent inference using the truncated sample, i.e. disregarding the units with missing data.

In the following I describe three estimation procedures, which relax the MCAR versus the MAR condition and correct for the selection on observables. In Section 2.3.1 I consider the inverse probability weighted estimator and the equivalent propensity score weighting estimation methods. These estimators relax the MCAR versus the MAR condition, without any cost in terms of additional assumptions to impose.

In Section 2.3.2 I present the imputation methods and I describe the conditions necessary for a correct inference of a conditional model of interest using the imputed data to replace the missing data. These conditions are stronger than the MAR assumption.

Finally, in Section 2.3.3 I show another possible solution for the selection on observables, which consists in correcting for the sample selection bias by introducing a control function, i.e. a correction term. Unfortunately this implies imposing, besides the MAR condition, some assumptions on the form of the control function.

2.3.1 Inverse probability weighted GMM estimator and the propensity score weighting methods

Let us consider again the GMM with moment condition function given by

$$E\left[\psi(y, x^y, x^c; \theta) \mid x^y, x^c\right] = 0, \tag{4}$$

holding in the absence of missing data. Let us suppose that the MAR is satisfied. Then, it is possible to control for selection on observables and to obtain a consistent GMM estimator by weighting the moment condition function by the inverse of the probability of selection given w, which is defined by Rosenbaum and Rubin (1984) as the propensity score, say $p(w) = Pr(r = 1 \mid w)$.

The new moment conditions for the weigthed GMM estimator become:

$$E\left[\psi(y, x^y, x^c; \theta)\frac{r}{p(x^y, x^c, x^r)} \mid x^y, x^c\right],$$

which, under the MAR condition, are equal to

$$E_{x^r}\left[E\left[\psi(y, x^y, x^c; \theta) \mid x^y, x^c, x^r, r = 1\right]\frac{Pr(r = 1 \mid w)}{p(w)}\right] = E_{x^r}\left[E\left(\psi(y, x^y, x^c; \theta) \mid x^y, x^c, x^r\right)\right] = 0$$

Note that the above weighted GMM is robust to any type of misspecification of the propensity score when the MAR and the IV exclusion restriction hold. This is because under the MAR and the IV exclusion restriction, we have

$$E\left[\psi(y, x^y, x^c; \theta) \mid x^y, x^c, x^r, r = 1\right] = E\left[\psi(y, x^y, x^c; \theta) \mid x^y, x^c\right].$$

Let us assume to use an incorrect propensity score given by $Pr(r = 1 \mid \tilde{w}) = p(\tilde{w})$, then the moment condition becomes

$$E\left(\psi(y, x^y, x^c; \theta) \mid x^y, x^c\right) E_{x^r}\left[\frac{Pr(r = 1 \mid w)}{p(\tilde{w})}\right] = 0.$$

Note also that when the explanatory variables used for the propensity score are the same explanatory variables used for the model of interest, say x, then either $r \perp\!\!\!\perp y \mid x$ and both the weighted GMM and the GMM ignoring the selection mechanism are consistent, or $r \perp\!\!\!\perp y \mid x$ is not valid and both the weighted GMM and the GMM ignoring the selection mechanism are inconsistent. Therefore using the weighted GMM is not worth in the absence of at least one explanatory variable in the selection model not considered in the model of interest.

The above weighted GMM estimation method is usually called propensity score weighting estimation. This estimation method has its roots in the inverse probability weighted estimator proposed by Horvits and Thompson (1952) to compute the population mean to take account that the units

in a sample have different probabilities to belong to the sample and to be respondent. This idea has been recently reconsidered by Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995) and Abowd, Crépon and Kramarz (1997) for the estimation of conditional means in the presence of missing data and by Rosembaum (1987), Imbens (2000) and Hirano, Imbens and Ridder (2002) for the evaluation of treatment effects. For a detailed presentation of the inverse probability weighting theory I refer to Woolridge (2002).

Both the Horvits and Thompson inverse probability weighted estimator and the propensity score weighting methods use the inverse propensity score to weight the units of the truncated sample when computing the sample counterpart of a population moment condition of interest. In both cases the estimation proceeds in two steps: first the estimation of the propensity score using a proper binary model, second the weighted GMM estimation of the parameters of interest by using the inverse propensity score as weight. The propensity score in the second step is substituted with its estimate computed in the first step. This obviously affects the estimator, which should be corrected to take account of the substitution of the true values of the propensity score with their estimates. Following Newey and McFadden (1994) it is possible to correct the estimator just by considering a generalized method of moment estimation, which uses the above moment condition together with a moment condition derived form the estimation of the propensity score, i.e. estimating jointly the selection model and the model of interest. Examples of application of this method, can be found in Abowd, Crépon and Kramarz (1997) and Inkmann (2002). The additional moment condition for the selection mechanism specified as a latent index model (for example as a probit or a logit model) is simply the score moment condition.

The propensity score weighting method can be easily extended to consider more complex selection models, for example to a multinomial model for the evaluation of the multi-treatments effects (see for example Imbens 1999) or to a discrete hazard model for the in the estimation of panel model in the presence of attrition (see for example Abowd, Crépon and Kramarz 1997 and Inkmann 2002).

2.3.2 The imputation methods

The imputation methods are usually extensively used to solve the missing data problem in sample survey data. The basic idea is to substitute the missing values with values computed using the observed variables.

Different methods have been proposed to compute the imputed values. A first distinction is between the donor and the model imputation methods.

Donor imputation methods substitute to a missing variable for a nonrespondent the observed variable for a respondent. There are different methods to match the nonrespondent with a responding donor. A reasonable criterion is to choose a donor with characteristics similar to the nonrespondent. Given a set of auxiliary variables w observed for both respondents and nonrespondents, then the matching can be based on a distance measure of these variables between units. The donor may be identified as the unit with observed variables w as close as possible to the nonrespondent. This is the nearest neighbour matching method. An alternative method is instead the near neighbour one, which defines a group of potential donors by selecting all respondents with distance from the nonrespondent lower or equal to a fixed threshold. The method consists then in imputing the group average or the observed value for a donor randomly selected from the group of potential donors to the nonrespondent.

Model imputation is instead based on a model to predict the unobserved variable as a function of a set of covariates observed for both respondents and nonrespondents. For example, it is possible to estimate a regression model using the sample of respondents and then use the estimated parameters to impute the predicted conditional average to the nonrespondents.

A method which is mixture between the two above methods is the predicted matching mean imputation (PMM). The PMM uses a linear regression model to predict the missing variable, say y, given a set of observable covariates w (auxiliary variables). The respondents are then divided into classes on the basis of the predicted value of y, and each nonrespondent is associated with the class with the closest mean predicted value of y. Finally, a randomly drawn respondent from the matched class is used as donor for the nonrespondent.

The donor imputation methods have the advantage to impute values that are always in the range of possible realisations. The model imputation methods can instead produce values, which lie outside the sampling space. On the other side the model imputation methods provide an easy and reasonable way to deal with a large set of auxiliary variables w containing both discrete and continuous variables.

Let y be the variable to be imputed, r be the dummy indicating the response w be the auxiliary variables used for the imputation procedure, and let the MAR assumption, $f(y \mid w, r) = f(y \mid w)$, be valid. Then the missing data problem may be solved by controlling for the set of variable w. This implies either a stratification of the sample based on the variables w, as in the donor imputation, or a conditional model to explain y as a function of the observed w, as in the model imputation. Rosenbaum and Rubin (1983) show that, when the above MAR condition is valid, it is sufficient

to control for the propensity score rather than the full set of variables w. The propensity score is the probability to respond given the variables w, i.e. $Pr(r = 1 \mid w)$.

Let us assume that $Pr(r = 1 \mid w) = p(z)$, then

$$Pr(r = 1 \mid y, p(z)) = E_z[Pr(r = 1 \mid y, z) \mid y, p(z)] = E_z[p(z) \mid y, p(z)] = p(z),$$

which implies $(y \perp\!\!\!\perp r \mid p(z))$ so that

$$E(y \mid r = 1, p(z)) = E(y \mid p(z)).$$

Using the truncated sample of units for which y is observed, we can estimate correctly predict y for a nonresponding unit by matching on the propensity score. Using this type of imputation becomes equivalent to applying the propensity score matching methods.

The idea behind the propensity score matching method is to match each nonresponding unit (unit for which y is not observable) with one or more responding units with a close propensity score, and to impute to each nonresponding unit the average y observed for the matched observed units. There are different methods of matching. Among these there are the nearest neighbour matching, the radius matching, the kernel matching. The nearest neighbour matching imputes to each nonresponding unit the observed value of y for the responding unit with the closest propensity score. The radius matching matches each nonrespondent with all the respondents with a propensity score whose distance (absolute difference) from the observed nonrespondent one is lower than a fixed threshold. The kernel matching uses all observations on y for the respondents to impute a value for each nonrespondent. The imputed value is computed as a weighted average of the observed y with weights as bigger as lower is the distance between the propensity scores observed for the nonrespondent and for the matched respondent. The weights are computed by using a kernel function and are standardised to sum to 1.

Imputation methods can be further distinguished following other criterions, in particular

- hotdeck versus colddeck imputation,
- deterministic versus stochastic imputation,
- single versus multiple imputation.

The difference between hotdeck and colddeck methods is the use of information from the same dataset rather than information from external datasets or other waves in the case of panel surveys. The stochastic methods add an error term to the predicted value computed using a conditional

model. Finally the multiple imputation methods impute several values instead of a single one to the missing variable y. These methods have been proposed to estimate consistently both the parameters of interest and the variance estimator (see for example Rubin 1989 and 1996).

The imputation procedures are generally carried out by the institutes conducting the surveys, so that their main aim is to produce consistent estimation of the population mean and total of certain variables and are not related to the estimation of specific regression models of interest. The advantages of the imputation methods are mainly two:

- to provide to the data-users a data set ready to be utilised without wondering about the nonresponse problem,
- to impute data possibly using all information on the data collection process and the sampling scheme even variables which are not included in the user-dataset.

The drawbacks consist instead in the potential bias in the inference when the MAR condition is not valid and/or when the imputation procedure is not adequate.

Much of the literature about imputation has focused attention on the underestimation problem for the variance of the estimates computed using imputed values (see for example Rubin 1989 and 1996) or in the possible bias in the estimation of totals, means or other simple statistics of the variable affected by missing (see for example Lessler and Kalsbeek 1992). In the rest of this section I instead focus attention on the potential inconsistency of the estimation of a general conditional model of interest by using imputed values. In particular, I define a set of conditions under which the imputation procedure can be used to produce consistent estimation using a GMM estimator, which would be consistent in absence of the sample selection problem.

Let y^I be the imputed value computed by using a donor or model imputation procedure with auxiliary variables $w = (x^y, x^c, x^r)$, which are observed for both respondents and nonrespondents. Let the MAR condition be valid and let the imputation procedure be consistent for the estimation of the conidtional mean $E(y \mid w)$. Then, it is possible to write:

$$E(y \mid w, r = 1) = E(y \mid w, r = 0) = E(y^I \mid w, r = 0) = E(y \mid w).$$

Let us consider again the GMM estimator introduce in Section 2.1 based on the moment conditions

$$E\left[\psi(y, x^y, x^c; \theta) \mid x^y, x^c\right] = 0, \tag{5}$$

and let us replace the missing values with the imputed ones, so that the new moments are

$$E\left[\psi(y^I, x^y, x^c; \theta)(1 - r) + \psi(y, x^y, x^c; \theta)r \mid x^y, x^c\right]. \tag{6}$$

Then, we can write

$$E_{x^r}\left[E\left[\psi(y^I, x^y, x^c; \theta) \,|\, w, r=0\right] Pr(r=0 \,|\, w) + E\left[\psi(y, x^y, x^c; \theta) \,|\, w, r=1\right] Pr(r=1 \,|\, w) \,|\, x^y, x^c, x^r\right]. \tag{7}$$

Since $E(y^I \,|\, w, r=0) = E(y \,|\, w)$ we have

$$E\left[\psi(y^I, x^y, x^c; \theta) \,|\, w, r=0\right] = \psi(E(y \,|\, w), x^y, x^c; \theta).$$

If $\psi(y, x^y, x^c; \theta)$ is a linear function in y,[4] then

$$E\left[\psi(y, x^y, x^c; \theta) \,|\, w, r=1\right] = \psi(E(y \,|\, w), x^y, x^c; \theta)$$

and the moments conditions with the imputed values can be rewritten as

$$E_{x^r}\left[E\left[\psi(y, x^y, x^c; \theta) \,|\, w\right]\right] = E\left[\psi(y, x^y, x^c; \theta) \,|\, x^y, x^c\right] = 0. \tag{8}$$

In conclusion, if the data are MAR, the model used to impute the data provides consistent estimation of $E(y \,|\, w)$, and the moment condition function is linear in y, then using the imputed values to replace the missing values implies a consistent GMM estimator. If the model to impute the missing y omits some of the explanatory variables relevant for the regression model of interest, for example the subset of variables x^y, then the GMM estimator using the imputed values may be not consistent.

If for example we consider the propensity score matching method to impute values for the missing y conditioning only on the propensity score, we can consistently estimate $E(y)$ but not $E(y \,|\, x^y, x^c)$, so that the GMM estimator with imputed data may be inconsistent.

2.3.3 GMM estimator corrected by using a control function

Another way to solve the problem of selection on observables is by correcting for the possible sample selection bias considering a control function (see Heckman and Robb 1985 and Heckman and Hotz 1989 for more details).

Let us consider again the GMM with moment condition function given by

$$E\left[\psi(y, x^y, x^c; \theta) \,|\, x^y, x^c\right] = 0, \tag{9}$$

holding in the absence of missing data. Let us suppose that the MAR is satisfied and that

$$E\left[\psi(y, x^y, x^c; \theta) \,|\, x^y, x^c, x^r\right] = \zeta(x^y, x^r, x^c),$$

[4] The likelihood score functions for a regression model with normal errors and for a latent index model for a binary variable are examples in which $\psi(.)$ is linear in y.

where ζ is a control function which form is supposed to be known. Then it is possible to solve the selection on observables by subtracting the above term from the moment function in the following way

$$E\left[\left(\psi(y, x^y, x^c; \theta) - \zeta(x^y, x^r, x^c)\right) r \mid x^y, x^c\right],$$

so that conditioning and then marginalizing with respect to x^r we obtain

$$E_{x^r}\left[E\left[\psi(y, x^y, x^c; \theta) - \zeta(x^y, x^r, x^c) \mid x^y, x^c, x^r\right] Pr(r = 1 \mid w)\right] = 0.$$

This estimation method, say GMM estimator corrected by using a control function, implies imposing additional assumptions on the form of the control function $\zeta(x^y, x^r, x^c)$, see Heckman and Hotz (1989) for an example in the case of a linear regression model.

Since both the imputation procedure and the control function $\zeta(x^y, x^r, x^c)$ require some additional assumptions, I suggest to solve the problem of selection on observables by using the weighted GMM with weights given by the inverse of the propensity score.

2.4 Selection on unobservable variables

Selection on unobervable variables occurs when the MAR condition does not hold. In this section I consider the methods relaxing the MAR condition and taking account of the consequent selection on unobsevables. In particular, in Section 2.4.1, I describe the parametric econometric selection model approach. The basic idea under the parametric econometric selection models is to specify jointly the model of interest together with the selection mechanism, allowing the errors in the two models to be correlated. In other words the parametric econometric selection models do not impose the MAR assumption, but specify a joint model for the dependent variable of interest y and the dummy indicating the selection r given a set of explanatory variables w, $f(y, r \mid w; \psi)$. The main critics to this approach concerns the restrictive assumption on the joint distribution of the errors, which is unfortunately untestable assumptions as well as the MAR is. In other words the parametric econometric selection model approach relaxes an untestable assumption by replacing it by another untestable assumption. The choice between either accepting the MAR condition or imposing a joint distributional assumption is not easy. Any decision is to some extent arbitrary and cannot be submitted to a test procedure. However, in Section 4, I will show an informal test to compare estimation procedures based on different untestable assumptions, which is based on the computation of bounds estimates instead of point estimates.

When the model of interest is a linear regression model, then it is possible to relax the joint distributional assumption imposed by the parametric selection model approach, by replacing it with

the additive separability condition, defined bellow, which is still an untestable assumption but it is weaker than the underlying assumptions of the parametric selection model approach.

Relaxing the parametric assumptions on the joint distribution of y and r requires the application of semiparamtric estimators. Among these estimators there are:

1. the Robinson's (1988) estimator,
2. the Powell's (1989) estimator,
3. the Cosslett's (1991) estimator.

In Section 2.4.2 I describe briefly those estimation methods, which belong to the semiparametic selection model approach. Moreover, I show that the Robinson's, the Cosslett's and the Powell's estimators are equivalent to the application of propensity score matching and stratification methods to the estimation of a linear regression model.

2.4.1 Parametric selection models

In the parametric selection model approach the selection mechanism is assumed to be a latent index model and quite often it is assumed to be a probit model. Let r be a binary variable, taking value 1 if a unit is observed and 0 if missing. Then it is supposed that r is related to a continuous latent variable r^* through the observation rule $r = 1\{r^* > 0\}$, where $1\{A\}$ is the indicator function of the event A, and the latent random variable r^* obeys the regression model

$$r^* = m_r(z;\gamma) + u_r,$$

where m_r is a non trivial function of the explanatory variables z, γ is a vector of parameters and the u_r is an error term identically and independently distributed (iid) with zero mean and unit variance[5] and independent from the variable z. In the case of a probit model the errors are assumed to be distributed as a Gaussian and $m_r(z;\gamma) = z\gamma$.

If the dependent variable of interest, y, is a dummy variable, then it is also assumed that it follows a latent index model, which it is often assumed to be a probit model. Then, the binary variable y is also related to a continuous latent variable y^* through the observation rule $y = 1\{y^* > 0\}$ and this latent random variable y^* obeys the regression model

$$y^* = m_y(x;\alpha,\beta),$$

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

where m_y is a non trivial function of the explanatory variables x, β is the parameter vector of interest, α is the intercept and u_y are iid errors distributed independently of x and z with zero mean and unit variance. In the case of a probit model $mm_y(x;\alpha,\beta) = \alpha + x\beta + u_y$, and the errors are normally distributed.

In the case of a continuous dependent variable, y^* is observed instead of the dummy y and the same type of model of y^* continue to hold.

Econometric selection models allow for the correlation between the error terms u_y and u_r so that the joint model becomes a censored bivariate model. In the case of normal error and of a binary model of interest we have a censore bivariate probit model with log-likelihood given by:

$$L^C = yr \ln \Phi_2(-\alpha - x\beta, -z\gamma; \rho) + (1-y)r \ln \Phi_2(\alpha + x\beta, -z\gamma; \rho) + (1-r) \ln \Phi(z\gamma), \tag{10}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standardized Gaussian distribution and $\Phi_2(\cdot,\cdot;\rho)$ denotes the cumulative distribution function of a bivariate Gaussian distribution with zero means, unit variances and correlation coefficient ρ. Maximising this censored likelihood it is possible to estimate the parameters β of the conditional model of interest.

If the dependent variable of interest is instead equal to the continuous variable y and assuming the following joint normality distribution for (y, r^*),

$$\begin{pmatrix} y \\ r^* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \alpha + x\beta \\ z\gamma \end{pmatrix} \quad \begin{pmatrix} \sigma_y^2 & \sigma_{yr} \\ & 1 \end{pmatrix} \right]; \tag{11}$$

then the log-likelihood for the censored joint model becomes:

$$L^C = r \ln \left[\phi\left(\alpha + x\beta; \sigma_y\right) \Phi\left(\frac{-z\gamma - \sigma_y \rho u_y}{(1-\rho^2)^{1/2}} \right) \right] + (1-r) ln \Phi\left(z\gamma\right), \tag{12}$$

where $\phi(\alpha + x\beta; \sigma_y)$ is the normal density with mean $\alpha + x\beta$ and variance σ_y^2, and $\rho = \sigma_{ry}/\sigma_y$ is again the correlation between the errors u_y and u_r.

The censored joint model to take account of missing data can be easily extended to other types of distribution of y and r. Given a generic density or probability distribution for y, $f(y \mid x;\theta)$, and a generic probability distribution for r

$$f(r \mid y, z) = Pr(r = 1 \mid y, z)^r Pr(r = 0 \mid y, z)^{1-r},$$

the censored log-likelihood becomes

$$L^C = r \ln [f(y \mid x;\theta) Pr(r = 1 \mid y, z)] + (1-r) ln Pr(r = 1 \mid z). \tag{13}$$

Since the joint distribution of (y, r) is censored the parameters of interest cannot be identified but when some IV exclusion restrictions are imposed. Therefore, the parametric selection models implicitly impose some IV exclusion restrictions. Summarising the underlying assumptions of the parametric selection models are:

1. a latent index model for the selection mechanism,
2. a specific joint parametric distribution for the errors of the model of interest and of the selection mechanism,
3. the independence between each of the two error terms and all the explanatory variables,
4. some IV exclusion restriction.

By joint distribution assumption I mean henceforth the above assumptions (1)-(4).

A well known and easier way, which is used to estimate continuous regression models in the presence of missing data, is the two-step Heckman (1979) procedure. This estimation procedure is based on the assumption that (y, r^*) are jointly normally distributed as in (11). Under the normality assumption, it is easy to prove that

$$\begin{aligned} E(y \mid x, z, r = 1) &= \alpha + x\beta + E(u_y \mid u_r > -z\gamma) \\ &= \alpha + x\beta + \rho\sigma_y E(u_r \mid u_r > -z\gamma) \\ &= \alpha + x\beta + \eta\frac{\phi(-z\gamma)}{\Phi(-z\gamma)} \\ &= \alpha + x\beta + \eta\lambda(z), \end{aligned} \tag{14}$$

where λ is the inverse Mill's ratio and $g(x) = \eta\lambda(z)$.

Then the equation of the dependent variable y can be written as follows:

$$y = \alpha + x\beta + \eta\lambda(z) + u \tag{15}$$

The parameters in the equation (15) can be consistently estimated by considering a new regression with an additional explanatory variable, which is the inverse Mill's ratio, $\lambda(z)$. In other words, the missing data problem reduces to a problem of variable omission. Then, the Heckman estimation procedure consists of two steps. The first step implies the estimation of a selection mechanism specified as a probit model. The second step involves the estimation of the main equation with an additional explanatory variable given by the inverse Mill's ratio, $\lambda(z)$, with parameter γ estimated in the first step. Since the error term in the new regression is heteroskedastic, a proper estimation should be used to produce consistent estimates of the standard errors of the estimated parameters.

Though easy to implement, the Heckman procedure has some limits because of its restrictive distributional assumptions, which are the same assumptions imposed by the ML estimation of censored joint parametric model. For this reason several extensions of the Heckman two-step procedure have been proposed (see Vella 1998 for a survey of these methods).

Lee (1983) has introduced an extension to the case of non-gaussian errors for the selection mechanism. He considers a generic distribution F for the error term u_r and he modifies it in a standard normal variable by applying the transformation $u_r^* = \Phi^{-1}(F(u_r)) = J(u_r)$. Then he consider a bivariate distribution for (u_y, u_r) given by

$$H(u_y, u_r) = \Phi_2(\frac{u_y}{\sigma_y}, J(u_r); \rho),$$

where Φ_2 is bivariate normal distribution with unit variances, 0 means and correlation ρ and the marginal distributions remain $\Phi(u_y)$ and $F(u_r)$. Considering the above bivariate normal distribution we have:

$$E(u_r \mid x, z, r = 1) = \frac{\phi(J(z\gamma))}{F(z\gamma)} = \frac{\phi(J(z\gamma))}{\Phi\left(J(z\gamma)\right)}.$$

Then the correction term to be added as an additional variable in the equation of interest is a modified inverse Mill's ratio, say λ^m, given by

$$\lambda^m = \frac{\phi(J(z\gamma))}{F(z\gamma)} = \frac{\phi(J(z\gamma))}{\Phi\left(J(z\gamma)\right)}.$$

Other extensions of the two-step Heckman procedure have been proposed to avoid any type of assumption on the error distribution for the selection mechanism.

Several semiparametric estimations have been then proposed, some of which are described in next section.

Furthermore, extensions of the Heckman procedure to consider panel model with individual effects have been proposed by Hausmman and Wise (1990), Verbeek and Nijman (1992), Woolridge (1995), Kyriazidou (1997) and Vella and Verbeek (1999). Extentions of the maximum likelihood to consider random individual effects correlated between the main equation and the selection process have also been considered see for example the paper of Jensen, Rosholm and Verner (2002), which compare different estimators for panel data with sample selection by a Monte Carlo simulation exercise.

2.4.2 Semiparametric selection models

Let us consider the following regression and selection models:

$$y = \alpha + x\beta + u_y, \tag{16}$$

$$Pr(r = 1 \mid w) = Pr(r = 1 \mid z) = Pr(1(\eta(u_r, z; \gamma) > 0)) = p(z; \gamma) \tag{17}$$

where

- u_y are iid with mean 0 and variance σ_y and independent of explanatory variables $w = (x, z)$,
- β are the parameters of interest, while γ are the nuissance parameters of the selection model,
- the propensity score $p(z; \gamma)$ is a nonlinear function from $\mathcal{Z}$ to the set of the real numbers, which does not lie in the space spanned by the variables x,
- u_r are iid with mean 0 and variance 1 and independent of explanatory variables z.

Conditioning the regression equation on the truncated sample, $r = 1$, and on the explanatory variables x of the regression model and z of the selection model, we have

$$E(y \mid r = 1, x, z) = \alpha + x\beta + E(u_y \mid r = 1, x, z).$$

We define the *separability condition*, imposed by the semiparametric selection models, as the following independence condition

$$(u_y \perp\!\!\!\perp (z, x) \mid p(z; \gamma), r). \tag{18}$$

This condition allows to write the regression equation for the truncated sample as

$$y = \alpha + x\beta + E(u_y \mid r = 1, p(z; \gamma)) + u, \tag{19}$$

where u is a residual error term with mean zero. Since $E(u_y \mid r = 1, p(z; \gamma))$ depends only on the propensity score, we can rewrite the equation (19) for the truncated sample as:

$$y = \alpha + x\beta + h(p(z; \gamma)) + u. \tag{20}$$

The separability condition can be alternatively defined as the existence of a nonlinear function $g(z)$ from $\mathcal{Z}$ to the set of the real numbers, which does not lie in the space spanned by the variables x such that

$$(u_y \perp\!\!\!\perp (z, x) \mid g(z), r).$$

Note that this definition of separability is equivalent to the existence of a function $g(z)$ which controls the selection bias as defined in Angrsit (1997).

In conclusion, the separability condition allows writing the regression equation for the truncated sample as

$$y = \alpha + x\beta + s(z) + u. \tag{21}$$

In the following I prove the equivalence between the condition

$$(u_y \perp\!\!\!\perp (z,x) \,|\, p(z;\gamma), r)$$

and

$$(u_y \perp\!\!\!\perp (z,x) \,|\, g(z), r).$$

The proof that

$$(u_y \perp\!\!\!\perp (z,x) \,|\, p(z;\gamma), r)$$

implies the existence of a function $g(z)$ such that

$$(u_y \perp\!\!\!\perp (z,x) \,|\, g(z), r)$$

is trivial as long as we choose $g(z) = p(z;\gamma)$. The proof that if there exists a function $g(z)$ such that

$$(u_y \perp\!\!\!\perp (z,x) \,|\, g(z), r = 1)$$

then

$$(u_y \perp\!\!\!\perp (z,x) \,|\, p(z;\gamma), r)$$

is equivalent to prove that any function $g(z)$ such that

$$(u_y \perp\!\!\!\perp (z,x) \,|\, g(z), r)$$

must be a function of the propensity score.

Note that

$$((u_y, r) \perp\!\!\!\perp w \,|\, p(z;\gamma))$$

is equivalent to

$$(u_y \perp\!\!\!\perp w \,|\, p(z;\gamma)), \quad (r \perp\!\!\!\perp w \,|\, u_y, p(z;\gamma))$$

or to

$$(u_y \perp\!\!\!\perp w \,|\, p(z;\gamma), r), \quad (r \perp\!\!\!\perp w \,|\, p(z;\gamma)).$$

Since $Pr(r = 1 \,|\, w) = p(z;\gamma)$ and $(u_y \perp\!\!\!\perp w)$, the above conditions

$$(r \perp\!\!\!\perp w \,|\, p(z;\gamma))$$

and

$$(u_y \perp\!\!\!\perp w \,|\, p(z;\gamma))$$

are always satisfied. As a consequence

$$(u_y \perp\!\!\!\perp w \,|\, g(z), r)$$

holds if and only if

$$(r \perp\!\!\!\perp w \,|\, u_y, g(z))$$

. Exploiting this last condition, we have

$$\int_{-\infty}^{+\infty} Pr(r = 1 \,|\, g(z), u_y) g(u_y \,|\, g(z)) du_y = Pr(r = 1 \,|\, g(z), w) = b(g(z)) = p(z; \gamma),$$

so that $g(z) = b^{-1}(p(z; \gamma))$, that is g(z) is a function of the propensity score.

Finally, the separability condition can be alternatively defined as:

$$(u_y \perp\!\!\!\perp (z, x) \,|\, p(z; \gamma), r = 1).$$

Given $u_y \perp\!\!\!\perp w$ we have that

$$u_y \perp\!\!\!\perp w \,|\, p(z; \gamma).$$

Then, by applying the law of the total probability to the density

$$f(u_y \,|\, p(z; \gamma), w) = f(u_y \,|\, p(z; \gamma))$$

we obtain:

$$f(u_y \,|\, p(z; \gamma)) = f(u_y \,|\, p(z; \gamma), r = 1) Pr(r = 1 \,|\, p(z; \gamma)) + f(u_y \,|\, p(z; \gamma), w, r = 0) Pr(r = 0 \,|\, p(z; \gamma)).$$

This is because by definition of the propensity score we have

$$Pr(r = 1 \,|\, p(z; \gamma), x, z) = Pr(r = 1 \,|\, p(z; \gamma)) = p(z; \gamma)$$

and

$$Pr(r = 0 \,|\, p(z; \gamma), x, z) = Pr(r = 0 \,|\, p(z; \gamma)) = 1 - p(z; \gamma).$$

Thence $f(u_y \,|\, p(z; \gamma), z, x, r = 0) = f(u_y \,|\, p(z; \gamma), r = 0)$ and $(u_y \perp\!\!\!\perp (z, x) \,|\, p(z; \gamma), r)$.

Often the semiparametric model assume that the selection model be a function of a single index linear in z, $v = z\gamma$, i.e. they assume that:

$$Pr(r = 1 \,|\, w) = Pr(r = 1 \,|\, z\gamma) = Pr(1(\eta(u_r, z\gamma) > 0)) = p(z\gamma).$$

Different methods have been proposed to estimate semiparametrically the coefficients of interest β. These methods consist of two following steps:

- estimating nonparametrically or semiparametrically the propensity score, $p(z;\gamma)$,
- estimating the equation (21) by controlling for the bias correction term $s(z)$.

To control for the bias correction term there are three main types of procedure:

1. applying to the equation (21) a deviation from the mean transformation with mean computed conditioning to $p(z;\gamma)$ or to $v = x\gamma$ estimated in the first step,
2. rewriting the equation (21) as difference between all possible pair of individuals in the sample and applying weights decreasing to 0 as the difference of $p(z;\gamma)$ or $v = x\gamma$ between a pair of individuals increases,
3. substituting the correction term, $s(z)$ by considering a set of dummy variables indicating the subsets of a partition of either the [0,1] support of the propensity score or the support of the variable $v = x\gamma$.

If the correction term $g(z)$ includes a constant in its definition, then it is not possible to identify the intercept of the equation of interest. This is beacause applying the deviation from the mean transformation in (1), the difference transformation in (2) or introducing a set of dummies as in (3) the intercept cancels out of the equation of interest. I will consider more extensively the intercept identification problem in the Section 2.4.2.

It is interesting to note that the tricks used to get rid of the correction term are similar to some estimation methods applied in panel data analysis. The method (1) is analogous to a within group estimator applied to get rid of the individual effect in a panel data model. The method (2) is similar instead to the Kyriazidou (1997) estimator for panel data with fixed effect and sample selection bias. Finally, the method (3) is analogous to estimation of a panel model with individual fixed effects, where the individual effects are replaced with effects associated to different levels of the propensity score.

In the following I present the Robinson's (1988), the Powell's (1989) and the Cosslett's (1991) estimators, which are semiparametric estimation procedures corresponding to the application of the above methods (1), (2) and (3) and using information on $v = z\gamma$ estimated in a first step. When instead information on the propensity score is used in the second step, the above methods (1), (2) and (3) are equivalent to the application of the propensity score matching and stratification methods for a linear regression model. As we will see later, the propensity score methods used to

estimate a regression model do not require the MAR condition, but they require the separability condition.

Propensity score methods have been recently considered in econometric by Heckman, Ichimura and Todd (1997), Heckman, Ichimura, Smith and Todd (1997), Dehejia and Wahba (1999, 2002) and Lechner (1999) and have been extended to the multi-valued treatment case by Imbens (2000). In the following we consider them in the special case of a regression model with missing data.

Robinson's estimator and propensity score matching methods. Let us consider the equation (19)

$$y = \alpha + x\beta + E(u_y \,|\, p(z;\gamma), r = 1) + u = \alpha + x\beta + h(p(z;\gamma)) + u,$$

and let us assume that the selection model be a function of a single index linear in z, $v = z\gamma$, i.e. $p(z;\gamma) = p(z\gamma) = p(v)$. Let us take the conditional expectation of the above equation with respect to v

$$E(y \,|\, v) = E(x \,|\, v)\beta + g(v).$$

By subtracting the terms in the last equation from the former one, we have

$$y - E(y \,|\, v, r = 1) = [x - E(x \,|\, v, r = 1)]\,\beta + u. \tag{22}$$

Estimating nonparametrically the expectation terms $E(y \,|\, v, r = 1)$ and $E(x \,|\, v, r = 1)$, we can estimate β by a simple regression. Robinson (1988) has proposed this type of estimator without specific reference to the econometric selection models. Its relationship with semiparametric estimation methods for sample selection problem (such as the Powell 1989 estimator) has been emphasised in Pagan and Ullah (1999), to which I refer for more details.

A possible solution to estimate nonparametrically $E(y \,|\, v, r = 1)$ is

$$\hat{E}(y_i \,|\, z_i\gamma, r = 1) = \hat{E}(y_i \,|\, v_i, r = 1) = \frac{\sum_{j=1}^{n} I(\,|\, v_i - v_j \,|\, \leq h/2) y_j}{\sum_{j=1}^{n} I(\,|\, v_i - v_j \,|\, \leq h/2)},$$

where h is a positive constant that goes to 0 for n which tends to infinity.

If we use the propensity score instead of v, the above estimation of $E(y \,|\, v)$ becomes a matching propensity score estimation.[6]

Note that the Robinson's estimator is similar to the two-step Heckman estimator. Considering again the equation (15) of the last section, which corrects for the selection bias by adding the

[6] See Section 2.3.2 for more details on the matching propensity score method.

inverse Mill's ratio among the explanatory variables,

$$y = \alpha + x\beta + \eta\lambda(z) + u, \tag{23}$$

and taking the conditional expectation with respect to $\lambda(z)$ we get

$$E(y \,|\, \lambda(z), r = 1) = E(x \,|\, \lambda(z), r = 1)\beta + \eta\lambda(z).$$

Then, subtracting the terms in the last equation from the previous one, we have

$$y - E(y \,|\, \lambda(z)) = [x - E(x \,|\, \lambda(z))]\,\beta + u. \tag{24}$$

From the above equation we can estimate β by regressing $[y - E(y \,|\, \lambda(z))]$ on $[x - E(x \,|\, \lambda(z))]$, where the conditional expectations are replaced by the ordinary least squares predictors. It is easy to prove that this estimation and the Heckman estimation of β give the same results. This is known as the equivalence of Frisch and Waugh (1933).

The main difference between the Robinson's estimator and the Heckman estimator is that the former does not impose restrictions on the functional form of $g(z)$, while the Heckman estimator specifies $g(z)$ as the inverse Mill's ratio.

Powell's estimator and propensity score matching methods. Let us consider the equation (19)

$$y = \alpha + x\beta + E(u_y \,|\, p(z;\gamma), r = 1) + u,$$

To eliminate from equation (19) the nuisance correction term $E(u_y \,|\, r = 1, p(z;\gamma))$, we can consider the difference between two generic units i and j, i.e.

$$y_i - y_j = (x_i - x_j)\beta + E(u_{y,i} \,|\, r_i = 1, p(z_i;\gamma)) - E(u_{y,j} \,|\, r_j = 1, p(z_j;\gamma)) + u_i - u_j,$$

and we can estimate β by a weighted least squares estimator with weights, w_{ij}, as smaller as higher is the difference between $p(z_i;\gamma)$ and $p(z_j;\gamma)$. This type of estimation procedure can be viewed as a propensity score matching applied to a linear regression model. If we assume that the selection model be a function of a single index linear in z, $v = z\gamma$, then the weights can be computed by considering the difference in v instead of the difference in the propensity score. A possible reasonable choice for the weights is then:

$$w_{ij} = \frac{I(\,|\, v_i - v_j \,|\, \leq h/2)}{\sum_{j=1}^{n} I(\,|\, v_i - v_j \,|\, \leq h/2)}, \tag{25}$$

where h is a parameter greater than 0. This second estimation procedure has been proposed by Powell (1989).

Cosslett's estimator and propensity score stratification methods. Let us consider again the equation (19)

$$y = \alpha + x\beta + E(u_y \,|\, p(z;\gamma), r = 1) + u,$$

then an alternative method to control for the correction term $E(u_y \,|\, r = 1, p(z;\gamma))$ is given by the propensity score stratification method.

The propensity score stratification method stratifies the sample in s disjoint sub-samples associated with s disjoint subintervals of the $[0,1]$ support of the propensity score, each one denoted by the index j, taking values from 1 to s, and then computes the average $E(u_y \,|\, r = 1, p(z;\gamma))$ by applying the law of total probabilty

$$E(u_y \,|\, p(z;\gamma), r = 1) = \sum_{p(z;\gamma)\in\{1,\ldots,s\}} E(u_y \,|\, p(z;\gamma) \in j, r = 1) I(p(z;\gamma) \in j).$$

For the truncated sample $E(u_y \,|\, p(z;\gamma) \in j, r = 1)$ depends only on j, so without loss of generality we impose $E(u_y \,|\, p(z;\gamma) \in j, r = 1) = \alpha_j$, where α_j is a constant unknown parameter. We can therefore correct for the $E(u_y \,|\, p(z;\gamma), r = 1)$ in the regression equation by adding s dummy variables indicating the subinterval to which the propensity score belongs, i.e.

$$y = x\beta + \sum_{p(z;\gamma)\in\{1,\ldots,s\}} \alpha_j I(p(z;\gamma) \in j) + u. \tag{26}$$

If we assume that the selection model be a function of a single index linear in z, $v = z\gamma$, then it is possible to stratify the sample by considering a partition of the support of $v = z\gamma$, instead of the support of the proepensity score. This method is followed by Cosslett (1991), who estimates the propensity score by a nonparametric method. The Cosslett's estimator is consistent but not asymptotically normal.

Note that we do not need the MAR assumption, thence $(u_y \perp\!\!\!\perp r \,|\, p(z;\gamma))$ to consistently estimate the β parameters. We need instead the condition $(u_y \perp\!\!\!\perp x, z \,|\, p(z;\gamma), r)$.

The β parameters can be equivalently estimated by applying to the regression equation (26) a deviation from the mean transformation, with mean computed conditioning to the set of dummy variables associated to the subintervals, and then by computing the OLS estimator (see the equivalence of Frisch and Waugh 1993). The Robinson's estimator is therefore equivalent to the Cosslett's estimator when using the above deviation from the mean transformation.

In the application I use the propensity score stratification method to compute a naive estimator to correct for the sample selection bias in a probit model of interest. I simply add into the probit

model a set of dummy variables and I allow for the variance of the error term to depend on the same set of dummies. The dummies are the indicators of the subintervals to which the propensity score belongs. 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. ,[7] Obviously the residual error term u may differ from a normal error, so that the consistency of this estimator in the case of a probit model is not ensured. 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 correcting for the possible heteroskedasticity of the error term by allowing its variance, σ, to depend on the above dummies in the following way:

$$\sigma = exp(\sum_{p(z;\gamma)\in\{1,...,s\}} \eta_j I(p(z;\gamma) \in j)),$$

may solve the possible ditributional misspecification.

Additional assumptions to estimate the intercept. The above semiparametric selection models estimators do not allow to estimate the intercept in the regression of interest. This is because the constant cancels out from the regression equation after applying filters such as difference transformations, $x_i - x_j$ as Powell does, deviatons from the mean, $[x_i - E(x_i \mid z_i\gamma), r = 1]$, as Robinson does, or by adding a set of dummy variables as in the Cosslett's estimator. Anyway, after the estimation of the slope coefficients, it is possible to compute the intercept by considering the observations for which $E(u_y \mid r = 1, x, z) = 0$, as suggested by Heckman (1990). To estimate the intercept, we can use the observations for which the propensity score is close to 1, i.e. the observations for which $E(u_y \mid r = 1, x, z) \simeq E(u_y \mid x, z) = 0$. Since the probability of selection into the sample increases monotonically with respect to $z\gamma$, it is possible to find the subset of z for which $Pr(r = 1 \mid z) > 1 - \epsilon$, say

$$\bar{Z} = \{z : p(z\gamma) > 1 - \epsilon\} = \{z : z\gamma > h\},$$

where $\epsilon \to 0$ and $h \to \infty$ for $N \to \infty$. Then, the intercept can be estimated as in a standard ordinary least squares estimation by computing the difference between the sample average of y

[7] The balancing property requires that the distribution of the variables z does not differ between respondents and nonrespondents given $p(z;\gamma) \in j$ for $j = 1, ..., s$. If the balancing score is not satisfied for a subinterval then I split the subinterval into disjoint smaller subintervals, until the properties is satisfied.

and the sample average of $x\beta$, but using only the information of the truncated sample for which $(z\gamma > h)$, i.e.

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(y_i - x_i\beta)I(z_i\gamma > h)}{\sum_{i=1}^{n} I(z_i\gamma > h)}.$$

Another possible estimation for the intercept has been proposed by Andrews and Schafgans (1996), who substitute to the indicator function $I(z_i\gamma > h)$ with a weighting function. It has been proved that these estimators are consistent and asymptotically normal if $z\gamma$ has sufficient mass function in the upper tail, i.e. if there exist enough values for the variables z such that $Pr(r = 1 \mid z)$ is close to 1.

2.5 Which are the costs to relax the MCAR, the MAR and the joint distributional assumptions?

This section describes the main findings concerning untestable and testable assumptions imposed by different estimation procedures in the presence of sample selection. A summary outline of these findings is shown in Table 7.

Relaxing the MCAR condition versus the MAR one does not have any cost when focusing attention on a conditional model of interest and estimating it with the inverse probability weighted GMM estimation or the analogous propensity score weighting GMM estimation method. This is because the MAR condition is nested into the MCAR condition and the consistency of the above estimator does not require additional assumptions. The MAR alone does not ensure instead the consistency of the GMM estimation for the truncated sample (i.e. ignoring the missing data), the GMM estimation with imputed data, the GMM corrected with a control function are used. Some additional assumptions must be imposed besides the MAR condition to solve the problem of the selection on observables for those estimators (see Table 7).

The GMM estimation using only the truncated sample is consistent if either the conditional MCAR condition or the MAR condition together to the IV exclusion restriction hold. The conditional MCAR is implied by the MCAR. The MCAR is indeed equivalent to require the MAR condition and $(r \perp\!\!\!\perp x^r \mid x^y, x^c)$, which is a testable condition. The IV exclusion restriction can be instead verified only if MAR holds.

The GMM estimation with imputed data is consistent if the MAR condition holds together with two additional assumptions: (i) the moment function must be linear in the dependent variable, y, (ii) the imputation procedure must provide a consistent estimation of $E(y \mid w)$. The first assumption

depends on the form of the model of interest, while the second depends on the type of imputation procedure adopted.

The GMM estimation corrected by using a control function is consistent if both the MAR condition and the assumptions of the form of the control function are satisfied.

Since the consistency of the inverse probability weighted GMM estimation requires only the MAR condition, this estimation method should be preferred to the others. Under the MAR, a Hausman type test may be used to verify the validity of the additional underlying assumptions imposed by alternative GMM estimators (the GMM estimators using imputed data, corrected with a control function and ignoring the missing data). However, note that the Hausman type test is a proper test only under the MAR condition, which is unfortunately untestable.

Note that the additional assumption required by the MCAR condition with respect to the MAR condition is instead testable. This additional assumption is $(r \perp\!\!\!\perp x^r, x^y, x^c)$ and can be verified because (r, x^r, x^y, x^c) are observed for all individuals.

Relaxing the MAR condition versus the joint distribution assumption has a cost which is not assessable. This is because both the MAR and the joint distribution assumption are not testable. Furthermore, the two assumptions are not nested, thence verifying one against the other is not possible. However, if the untestable joint distribution assumption imposed by the parametric selection models holds, then it is possible to verify the validity of the MAR condition.

To be more specific the joint distribution assumption, imposed by parametric selection models, is composed by a set of assumptions: the parametric form of the errors joint distribution, the IV exclusion restrictions, the independence between errors and explanatory variables and the latent index model assumed for the selection process (see Section 2.4.1). Since r and z are always observable, the assumption of a latent index selection model $Pr(r = 1 \mid z) = Pr(u_r > m_r(z; \gamma))$ is testable as well as the independence between u_r and z. The assumptions on the joint distribution of the errors, $f(u_y, u_r)$, and the independence between u_y and x and between u_y and x^r (the IV exclusion restrictions) are instead untestable.

Under the MAR condition the above untestable assumptions become testable, but the joint density distribution of the errors becomes trivially equal to the product of the marginal density, $f(u_y, u_r) = f(u_y)f(u_r)$, which implies that selection on unobservable is not allowed by assumption.

Relaxing the joint distribution assumption versus the separability assumtpion has at least two costs: the lack of identification of the intercept and the restriction of the attention to linear

regression models. The separability assumption is weaker than the joint distribution assumption imposed by the parametric selection models. Angrist (1997) proves that a subset of assumptions imposed by the parametric selection models - more precisely the latent index selection model assumption, the independence between the error and the explanatory variables in the equation of interest and the IV exclusion restrictions - implies the separability condition. Moreover, the latent index selection model assumption can be substituted with a weaker assumption of a monotonic selection model.

A selection model is monotonic if, given two values for the vector z, say z^1 and z^2, then either

$$Pr(r = 1 \mid z^1, u_y) > Pr(r = 1 \mid z^2, u_y) \text{ or}$$

$$Pr(r = 1 \mid z^2, u_y) > Pr(r = 1 \mid z^1, u_y)$$

is almost surely true. The monotonocity condition is satisfied by any latent index selection model with constant parameters and errors independent from the explanatory variables. Let us consider, for example, a probit model for the selection mechanism and a linear regression model with Gaussian errors independent of (x, z) and a correlation between the error terms in the two models equal to ρ, then

$$Pr(r = 1 \mid z, u_y) = Pr\left(u_r \leq \frac{z\gamma + \rho u_y}{(1-\rho)^{1/2}}\right) = \Phi\left(\frac{z\gamma + \rho u_y}{(1-\rho)^{1/2}}\right).$$

Given two values of the variables z, say z^1 and z^2, then either

$$\Phi\left(\frac{z^1\gamma + \rho u_y}{(1-\rho)^{1/2}}\right) \leq \Phi\left(\frac{z^2\gamma + \rho u_y}{(1-\rho)^{1/2}}\right)$$

or

$$\Phi\left(\frac{z^2\gamma + \rho u_y}{(1-\rho)^{1/2}}\right) \leq \Phi\left(\frac{z^1\gamma + \rho u_y}{(1-\rho)^{1/2}}\right)$$

is almost true. If the errors in the equation of interest are not normal, the monotonicity condition continues to hold. This is because it is always possible to decompose the error term in the probit model in the following way:

$$u_r = E(u_r \mid u_y) + \epsilon = m(u_y) + \epsilon$$

Since Angrist (1997) considers the same set of explanatory variables for the selection model (propensity score) and for the model of interest, he does not require the IV exclusion restrictions to prove the separability condition.

For this reason, in the following, I give a proposition slightly different from Angrist (1997), which fits with the following regression and selection models:

$$y = \alpha + x\beta + u_y, \tag{27}$$

$$Pr(r = 1 \mid x^c, x^y, x^r) = Pr(r = 1 \mid x^c, x^r) = Pr(r = 1 \mid z) = p(z;\gamma) \tag{28}$$

where u_y are iid with mean 0 and variance σ_y.

Proposition 1 *Let us consider the models (27) and (28) and let us assume that the three following conditions hold:*

1. *the independence of between error and explanatory variables in the equation of interest, i.e.* $u_y \perp\!\!\!\perp x$,
2. *the IV exclusion restrictions, i.e.* $u_y \perp\!\!\!\perp x^r$.

Then the separability condition $(u_y \perp\!\!\!\perp w \mid p(z;\gamma), r)$ *is satisifed if and only if* $(r \perp\!\!\!\perp w \mid u_y, p(z;\gamma))$.

Proof

The condition separability condition $(u_y \perp\!\!\!\perp w \mid p(z;\gamma), r)$ *together with the condition* $r \perp\!\!\!\perp w \mid p(z;\gamma)$, *which is satisfied because* $Pr(r = 1 \mid w) = p(z;\gamma)$, *are equivalent to the condition*

$$((u_y, r) \perp\!\!\!\perp w \mid p(z;\gamma)),$$

which holds if and only if the two following conditions are satisfied:

$$(u_y \perp\!\!\!\perp w \mid p(z;\gamma)) \text{and} (r \perp\!\!\!\perp w \mid u_y, p(z;\gamma)).$$

$(u_y \perp\!\!\!\perp w \mid p(z;\gamma))$ *is always satisfied because* $u_y \perp\!\!\!\perp w$ *and* $p(z;\gamma)$ *is a function of* w, *so that the separability condition is satisfied if and only if* $r \perp\!\!\!\perp (x, z) \mid u_y, p(z;\gamma)$.

Angrist (1997) (Proposition 3) proves that when the selection model is monotonic then

$$(r \perp\!\!\!\perp w \mid u_y, p(z;\gamma),$$

, so that the separability condition is satisfied.

A latent index model assumption, $Pr(r = 1 \mid w)) = Pr(u_r > m_r(z;\gamma))$, the independence of the error and the explanatory variables in the regression model, and the IV exclusion restrictions, $u_r \perp\!\!\!\perp x^r$, are therefore sufficient to ensure the separability condition.

Since the parametric selection models assume the latent index model assumption, the IV exclusion restrictions, and the independence of u_y from x, their underlying assumptions imply the assumptions of the semiparametric models. In other words it is proved that the parametric selection models are based on stronger assumptions than the semiparametric selection models.

Since the assumptions of the semiparametric selection methods are nested into the assumptions of the parametric selection ones and the latter estimator is more efficient when its underlying assumptions are valid; I suggest to use a Hausamn type test to verify the former set of assumptions against the latter ones by verifying the equality of the slope coefficients estimators.

Unfortunately there are situations in which the intercept is the main parameter of interest. In that case some additional assumptions are required to ensure the consistency of the intercept estimators proposed by Heckamn (1990) and Andrews and Schafgans (1996), see Section 2.4.2. With this additional conditions the underlying assumptions of the semiparametric methods are not nested in the assumptions imposed by the parametric methods, so that it is difficult to fix an order of preference.

3 Analysis of poverty

In this section I show the results of the estimation of the poverty of being poor for Italy. I use a probit model and I take account of the missing data problem by using the propensity score weighting, the econometric selection approach and the imputation procedures described in the previous sections.

I use 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. The analysis of poverty is carried out for Italy in 1998.

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). At the moment the ECHP data are available for the first 5 waves, 1994-1998. A more detailed description of the ECHP can be found in Peracchi (2002).

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 expressed in thousands of ECU at constant 1995 prices (I use the equivalized size, OECD modified scale), which, henceforth, I will call briefly household income.

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 4.1. 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 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. It is quite troublesome to consider the household nonresponse using the ECHP-UDB. I computed the percentage of households unit nonresponding

in 1998 and I found quite strange results. For example, it seems that in France, Greece, the Netherlands and the UK (BHPS) the percentage of household unit nonresponses is equal to 0. For this reason, I do not consider the household unit nonresponse and I correct for it by considering the weights available in the ECHP-UDB, which anyway do not change the estimation results.

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). 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

The most common model used in empirical works to describe the poverty probability as a function of the variables characterising 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. The aim of the empirical analysis is to compare different estimation procedures of a probit model for the poverty probability, which take account of the sample selection problem 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. Note that the probit assumption is untestable wihtout information on the missing income variables, Neverhteless, under the MAR condition, the assumption is testable and we find out that it is not rejectable. To verify the normality of the errors in a probit model, we use the score test for normality in an ordered probit, adequately modified to consider a binary dependent variable instead of an ordered categorical one, presented in Machin and Stewart (1990), which is in turn a modification of the score test derived by Chesher and Irish (1987) for a grouped dependent variable.

I apply six different types of estimations:

1. the probit with imputed data (imputation method), i.e. the estimation of a simple probit

model by replacing the unknown poverty dummy which the one computed using the imputed income,

2. the propensity score weighting method (weighing method), i.e. a probit with weights equal to the inverse probability of being respondent,

3. the censored bivariate probit (joint model), which models jointly the poverty probability and the response probability allowing the errors to be correlated,

4. the probit with complete data (ignoring method), i.e. excluding all individuals with a problem of nonresponse on the household income,

5. the propensity score stratification method (stratifcation method), which estimates a probit model corrected for the possible sample selection bias and for the possible heteroskedastic of the error term by considering dummies, which are indicators of disjoint subintervals to which the propensity score belongs,

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.

All estimation 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 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 labour 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:

- 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. Variables with such types of characteristics are used to solve the problem of identification in the censored bivariate probit, and they are called instrumental variables, IV.

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 rejectable, at least under the joint distributional assumption. Furthermore a Hausman type test to verify the equality of the weighted and the unweighted probit 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.

4 Comparison of inference methods to treat the missing data

There are several papers trying to assess and to compare different estimation methods to deal with the missing data problem or with the close problem of the evaluation of the causal effect of treatments from non-randomised experiments.

Among papers giving good surveys of the evaluation methods there are Angrist and Krueger (1999), Blundell and Costa-Dias (2002), Heckman, Ichimura, Smith and Todd (1997) and Heckman, Lalonde and Smith (1999).

Heckman, Ichimura, Smith and Todd (1997) use experimental data to verify the validity of the assumptions justifying the matching, the econometric selection models and the differences in differences estimation methods. They consider an experimental control group and a non-experimental comparison group, which do not receive the treatment, so that differences in the outcome variable between the two groups merely reflects the selection bias. Heckman, Lalonde and Smith (1999) study instead the sensitivity of alternative evaluation estimation methods by simulating different dataset assuming specific models for the response variable and for the treatment assignment mechanism. In both 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 perfor-

mance of the different estimations by verifying if these estimations, in my case the poverty probability estimation, lies outside the Manski bounds computed assuming very weak or no assumptions. In the following I describe the procedure adopted for this informal test.

4.1 Computation of the bounds

Following the approach used in Manski (1995), Horowitz and Manski (1998), Manski and Pepper (2000), Alvarez *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. The probability of poverty, $\Pr\{Y < c\}$, cannot be identified using only data on households with complete responses. By using the law of total probability, it is possible to decompose the probability of poverty as

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

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.

$$UB = \Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\} + \Pr\{r = 0\},$$
$$LB = \Pr\{Y < c \,|\, r = 1\} \Pr\{r = 1\}.$$

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 income subcompoenents are missing and other are observed, so that the hosuehold 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 value partially reported 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:

$$\Pr\{Y < c \,|\, r = 0\} = \Pr\{Y < c \,|\, Y^r < c, r = 0\} \Pr\{Y^r < c \,|\, r = 0\} + \\ + \Pr\{Y < c \,|\, Y^r \geq c, r = 0\} \Pr\{Y^r \geq c \,|\, r = 0\}.$$

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 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 us to compute the following new upper bound, called reported income upper bound,

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

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

Furthermore, I impose different types of weak assumptions that can help narrowing further the worst case 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.

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, 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

$$\begin{aligned} LB_{IV} &= \sup_z \Pr\{Y < c \,|\, x, z, r = 1\} \Pr\{r = 1 \,|\, x, z)\} \\ &\leq \Pr\{Y < c \,|\, x\} \\ &\leq \inf_z \Pr(Y < c \,|\, x, z, r = 1\} \Pr\{r = 1 \,|\, x, z\} + \Pr\{r = 0 \,|\, x, z\} \\ &= UB_{IV}. \end{aligned}$$

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 workers, then the bounds for $\Pr\{Y < c \,|\, x, z\}$, say MIV bounds, are given by

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

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.

I compute separately the IV bounds and the two MIV bounds conditioning to the set of above variables. 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,

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

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

The bounds computed using the reported income narrow down, 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 (see Section 2.3.2) may fail. In particular, the probit model for the probability of being poor implies a score function, which is linear in the dummy indicating the poverty status but is not linear in the household income, so that the consistency of the probit estimator is not ensured.

4.2 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 boostrap (1000 samples with replacement are drawn from the original data) and by taking the 5th percentile and the 95th percentile of the boostrap 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.

In Table 4 I report percentages of inconsistencies, i.e. percentages of cases in which the conditional predicted poverty probabilities lie outside the Manski bounds. To compute these inconsistencies, I substitute the lower (upper) bound with the lower (upper) confidence band estimated by bootstrap with 1000 replications and taking the 5th (95th) percentile of the corresponding distributions. 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 obviously is not adequate 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 that may affect this estimator when the IV used in the selection models are not very significant. The imputation method presents also some problems but in the case of a poverty line defined as 60% of the median income and using the IV. If we accept the MAR condition, then those problems may be due to the nonlinearity in the household income of the first order condition for the maximization of the probit likelihood or to a misspecification of the model used to impute the household income (see Section 2.3.2). The first order condition for the likelihood of a probit model is linear in the dummy indicating poverty, however it is the household income to be imputed in the ECHP and not the poverty status. Thence linearity in the household income is not satisfied.

Then, I investigate the consequences of omitting relevant variables (the number of workers and the dummies indicating the labour 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 stratification 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.

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.

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 selection models, 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 selection models may be seriously affected by the possible misspecification of the selection model, while the other methods seem to be more robust.

5 Conclusions

In this paper I have shown that even when the percentage of nonresponses is high, it is possible to narrow down the Manski bounds for the poverty probability using the information on partially reported income. In the empirical application the income information is missing in about 20% of cases, nonetheless using the partially reported income it is possible to limit the interval of possible values for the poverty probability from 20 to 7 percentage points.

Furthermore, I have shown that 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 quite well in detecting cases in which the estimators are inconsistent. In particular when misspecifying the poverty model and/or the selection model the estimators presents high percentage of cases in which the predicted poverty probabilities lie outside the Manski bounds.

References

Abowd J., Crépon B., Kramarz F. (1997), "Moment estimation with attrition", NBER working papers, 214.

Andrews D., Schafgans M. (1998), "Semiparametric estimation of the intercept of a sample selection model", *Review of Economic Studies*, vol. 65, 497-517.

Angrist J.D. (1997), "Conditional independence in sample selection models", *Economic Letters*, vol. 54, 103–112.

Angrist J.D., Imbens G.W., Rubin D.B. (1996), "Identification of causal effects using instrumental variables", *Journal of the American Statistical Association*, vol. 91, 434, 444–472.

Angrist J.D., Krueger A.B. (1998), "Empirical strategies in labor economics", Princeton University, Industrial Relation Section, Working Paper, 401.

Blundell R., Costa-Dias M. (2002), "Alternative approaches to evaluation in empricial microeconomics", IFS, Cemmap working paper, 10.

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.

Dehejia R.H., Wahba S. (1999), "Causal effects in nonexperimental studies: reevaluation of the evaluation of training programs", *Journal of the American Statistical Association*, vol. 94, 1053–1063.

Dehejia R.H., Wahba S. (2002), "Propensity score matching methods for non-experimental causal studies", *Review of Economics and Statistics*, vol. 84, 1, 151–161.

Engle R.F., Hendry D.F., Richard J.-F. (1983), "Exogeneity", *Econometrica*, vol. 51, 2, 277–304.

Eurostat (1995), "Cross-sectional imputation rules and application to the micro-data files", PAN 47.

Eurostat (1999), *ECHP UDB Manual, Waves 1,2 and 3.*

Eurostat (2000), "Imputation of income in the ECHP", PAN 164/00.

Frisch R., Waugh F. (1933), "Partial time regressions as compared with individual trends", *Econometrica*, vol. 45, 939–53.

Hansen L. (1982), "Sample Properties of Generalized Method of Moments Estimators", *Econometrica* 50, 1029–1054.

Hausman J.A., Wise D.A. (1979), "Attrition bias in experimetnal and panel data: the Gary income maintenance experiment", *Econometrica*, vol. 47, 455-473.

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

Heckman J. (1990), "Variaties of selection bias", *The American Economic Reiwew*, vol. 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*, vol. 84, 408, 862–874.

Heckman J., Ichimura H., Smith J., Todd P. (1998), " Characterizing selection bias using experimental data", NBER working papers, 6699.

Heckman J., Ichimura H., Todd P. (1997), "Matching as an econometric evaluation estimator: Evidence from Evaluating a Job Training Program", *Review of Economic Studies*, vol. 64, 4, 261-294.

Heckamn 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*, vol. 3, North Holland, Amsterdam.

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. (1998), "Combining panel data sets with attrition and refreshment samples, NBER working papers, 230.

Heckman J.J., Robb R. (1985), "Alternative methods for evaluating the impact of interventions", in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge.

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

Horowitz J.L., and Manski C.F. (2002), "Identification and estimation with incomplete data", mimeo.

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

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

Imbens G.W., Angrist J.D.(1994), "Identification and estimation of local average treatment effects", *Econometrica*, vol. 62, 2, 467–475.

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

Kyriazidou E. (1997), "Estimation of a panel data estimation model", *Econometrica*, vol. 65, 1335-1364.

Lechner M. (1999), "Earnings and employment effects of continuous off-the job training in East Germany after unification", *Journal of Business & Economic Statistics*, vol. 17, 1, 74-90.

Lee L.F. (1983), "Generalized linear model with selectivity", *Econometrica*, 51, 2, 507-512.

Lessler J.T., Kalsbeek W.D. (1992), *Nonsampling Error in Surveys*, John Wiley & Sons, New York.

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, pp 327-350.

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

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

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

Mátyás L (1999), *Generalized Method of Moments Estimation*, Cambridge University Press, New York.

Newey W., McFadden D. (1994), "Large sample estimation and hypothesis testing", in R. Engle and D. McFadden, (eds.), *Handbook of Econometrics*, vol. 4, 2111-2245, North Holland, Amsterdam.

Nicoletti C. (2002), "Non-response in dynamic panel data model", Working papers of the Institute for Social and Economic Research, paper 2002–31, Decembre, Colchester: University of Essex

Pagan A., Ullah A. (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

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

Powell J.L. (1989), "Semiparametric estimation of censored selection models", Department of Economics, University of Wisconsin-Madison.

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.

Robins J., Rotnitzky A. (1995), "Semiparametric efficiency in multivariate regression models with missing data", *Journal of the American Statistical Association*, vol. 90, 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*, vol. 90, 106-121.

Robinson (1988), "Root-N-consistent semiparametric regression", *Econometrica*, vol. 56, 931–954.

Rosenbaum P.R. (1987), "Model–based direct adjustment", *Journal of the American Statistical Association*, vol. 82, 387-394.

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

Rubin D.B. (1976), "Inference and missing data", *Biometrika*, vol. 63, 581–592.

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

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

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

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 Alvarez 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 Alvarez 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 Reosurces*, vol. 3, 127–169.

Vella F., Verbeek M. (1999), "Two-step estimation of panel data models with censored endogenous variables and selection biad", *Journal of Econometrics*, vol. 90, 239–263.

Verbeek M., Nijman T. (1992), "Incomplete panel and selection bias", in L. Mátyás , P. Sevestre (eds.), *The Econometrics of Panel Data*, Kl'uewer Academic Publishers, New York.

Vytlacil E. (2002), "Independence, monotonicity, and latent index models: and equivalence result", *Econometrica*, vol. 70, 1, 331-341.

Wooldridge J.M. (1995), "Selection correction for panel data models under conditional mean independence assumptions", *Journal of Econometrics*, vol. 68, 115-132.

Wooldridge J.M. (2002), "Inverse probability wighted M-estimators for sample selection, attrition and stratification", IFS, Cemmap working paper, 11.

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
Propensity score stratification	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	Stratification	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	Stratification	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	Stratification	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	Stratification	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	Stratification	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	Stratification	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	Stratification	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	Stratification	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	Stratification	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

Table 7: Summary outline of the underlying assumptions and costs of different estimators.

	Relaxing MCAR versus:	Relaxing MCAR versus:
Assumptions B	Conditional MAR	MAR, IV exclusion restrictions
Consequences	Focusing on the conditional model for y	Focusing on the conditional model for y
Estimation methods	GMM (ex. OLS, IV, ML)	GMM (ex. OLS, IV, ML)
Sample used	Truncated sample	Truncated sample
Cost	None (MAR is nested in MCAR)	Assumptions B except MAR (which is nested in MCAR)
	Relaxing MCAR versus:	Relaxing MCAR versus:
Assumptions B	MAR, moments linear in y and consitent imputation estimation	MAR and assumptions on the form of the selection bias
Consequences	Selection on observables	Selection on observables
Estimation methods	GMM using imputed data	GMM corrected by considering a control function estimator
Sample used	Full sample	Full or truncated sample
Cost	Assumptions B except MAR (which is nested in MCAR)	Assumptions B except MAR (which is nested in MCAR)
	Relaxing MCAR versus:	Relaxing MAR versus:
Assumptions B	MAR	Joint distribution
Consequences	Selection on observables	Selection on unobservables
Estimation methods	Inverse probabiltiy weighted GMM Propensity score weighting methods	ML of the joint model (paremetric seleciton model method)
Sample used	Full sample	Full sample
Cost	None (MAR is nested in MCAR)	Assumptions B (which are not nested in MAR)
	Relaxing joint distribution versus:	Relaxing joint distribution versus:
Assumptions B	Separability condition Heckman(1990) or Andrews and Schafgans (1996) assumptions	Separability condition
Consequences	Focusing on a linear regression model	Focusing on the slope coefficient of a linear regression for y
Estimation methods	Semiparametric estimation, propensity score stratification, matching methods	Semiparametric estimation, propensity score stratification, matching methods
Sample used	Full sample	Full sample
Cost	Assumptions B except separability (which is nested)	None (separability is nested in the joint distribution assumption)

In the consequences' cells I indicate only the consequences that add up to the ones of the previous cell.

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,

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

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" $f_h > 1$. The latter, "common for the whole household and all personal level income components in it, is introduced to correct for the effect of nonresponding individuals within an otherwise responding household in the construction of household level variables such as total net income .... All components reported at the personal level are multiplied by this factor" (Eurostat 2000, p. 8).

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 equivalised 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

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

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), [8] 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.

[8] 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 2000 and Raghunathan, Solenberger and Hoewyk 1999).

The IVE procedure allows to define 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.