Revista de Análisis Económico, Vol. 25, Nº 2, pp. 193-221 (Diciembre 2010)

INCOME NONRESPONSE AND INEQUALITY MEASUREMENT*

FALTA DE RESPUESTA EN INGRESOS Y MEDICION DE LA DESIGUALDAD

GUILLERMO PARAJE
School of Management, Universidad Adolfo Ibáñez

MELVYN WEEKS
Faculty of Economics and Politics, University of Cambridge

Abstract

This paper analyses the effects that income nonresponse has on certain well-known inequality coefficients (e.g. Gini, Theil and Atkinson indexes). A number of statistical methods have been developed to impute missing values of incomes for nonrespondents. By simulating several patterns of income nonresponse on actual sub-samples of the Argentinean household survey, this essay analyses the effects that different correction methods produce on a set of inequality coefficients. It is proved that methods often used to correct for nonresponse can introduce important biases on inequality coefficients if the patterns of missingness assumed by such methods do not coincide with the actual pattern.

Keywords: *Household surveys, income imputation, inequality measures.*

JEL Classification: *C15, C81, D31.*

Resumen

Al simular varios patrones de no-respuesta sobre muestras reales de la Encuesta Permanente de Hogares de Argentina, el artículo analiza los efectos que los métodos de imputación tienen sobre los indicadores

[*] We are much indebted to the helpful comments of Gabriel Palma, Bob Rowthorn, Carlos Rodríguez, Alvaro Angeriz, Paul Ryan, Andrew Kennedy and two anonymous referees. The usual caveat applies.

de desigualdad frecuentemente utilizados (e.g. Gini, Theil e índices de Atkinson). Se demuestra que métodos usualmente utilizados para corregir la falta de respuesta en ingresos pueden sesgar de manera importante a los indicadores de desigualdad, si los patrones de no-respuesta supuestos por dichos métodos no coinciden con los existentes en los datos reales.

Palabras Clave: *Encuestas de hogares, imputación de ingresos, medidas de desigualdad.*

Clasificación JEL: *C15, C81, D31.*

I. INTRODUCTION

The measurement of income inequality plays a pivotal role in the assessment of economic welfare and should play a central role in the design and implementation of social policies. A primary input for the measurement of inequality is income data coming from household surveys. It is a well-documented fact that in such surveys errors occur in the measurement of income (Atkinson *et al.*, 1995, for instance). Broadly speaking, two types of errors may exist: sampling errors, which are present when sampled households do not represent the population (because of a problem with the survey design itself or because a number of sampled households refuse to answer any question); and non-sampling errors, that are present when the information recorded for households is not accurate. In turn, two types of non-sampling errors may coexist. The first one is income underreporting, which occurs when people report incomes lower than the actual ones, while the other one is income nonresponse which is present whenever individuals refuse to answer income questions. While the former is associated with non-labour income, such as capital gains, rents and utilities, the latter is primarily associated with incomes from labour. The former can be relatively difficult to spot and correct (*e.g.* individuals earning rents cannot be distinguished from the rest, unless they declare that fact), while the latter is relatively easy to detect (individuals declare that they have worked but refuse to declare their earnings). The effect that income nonresponse has on inequality measurement is the central topic of this essay.

When a large proportion of individuals do not answer income questions the usefulness of household surveys as a tool to gather relevant information (to estimate inequality, for instance) is undermined. In many developed countries (*e.g.* Finland, Sweden) there are other sources of information apart from household surveys (such as administrative or tax records) that can be either used to measure inequality or to correct household surveys when they contain a significant number of individuals not responding about their incomes, so that inequality coefficients obtained from them reflect the true distributional situation. But in a large number of countries (mainly developing ones) such sources do not exist or are as unreliable as the household surveys. Administrative records, such as pension or retirement records, are incomplete, as only a portion of the population receive such benefits, whereas tax records are even less reliable, as tax evasion, especially on income taxes, is substantial. In most Latin American countries, for example, where income nonresponse rates are well above 5% (*e.g.* Chile, Colombia, Ecuador, Venezuela) and in certain cases are higher than 10% (*e.g.* Argentina, Costa Rica, Honduras, Panama),[1] no secondary sources are available to correct for such a high nonresponse. In these cases, and given the magnitude of this error, the first question that a survey analyst confronts, before any attempt to measure inequality is made, is what to do with individuals not responding to income questions.

The treatment of missing data in household surveys has been the main topic of an extensive literature, not only in economics but also in other social and non-social sciences (King *et al.*, 2001; Briggs *et al.*, 2003). In this literature, a number of statistical methods that correct for item (and, specifically, income) nonresponse have been suggested and used. Nevertheless, the specific question of how such methods affect inequality measurement has received significantly less (if any) consideration. Within the economic literature, comprehensive studies on household surveys, such as Deaton (1997), or on income inequality, such as Atkinson and Bourguignon (2000 a) and Silber (1999), pay little attention to issues related to the effect that data quality, in general, and income nonresponse, in particular, has on inequality inferences. Even though it is acknowledged that "observed monetary disposable income may give a biased representation of the actual income distribution of (monetary) income in a society" because of the existence of errors in the measurement of income (Atkinson and Bourguignon, 2000 b, p. 27) no reference is made to how such biases should be removed when surveys are the only source of information (which is the case of a large number of countries) to estimate inequality correctly. Nor do they try to quantify how biased inequality coefficients can be when the data is contaminated by income nonresponse.

This essay attempts to fill in this gap by analysing how inequality inferences can be affected by the use of different correction methods, under several patterns of income missingness (*i.e.* how nonrespondents are distributed across the income distribution). In this respect, the essay quantifies the biases that particular correction methods introduce in a number of well-known inequality coefficients (*e.g.* the Gini coefficient, the Theil and Atkinson indexes). Given the lack of secondary sources to infer the pattern of missingness in real data (a constraint usually faced in empirical studies), a simulation approach is used. The simulations consist in "contaminating" samples of workers that fully report their labour incomes from a particular survey, the Argentinean Permanent Household Survey. It is thus assumed that, following several patterns of nonresponse, a number of workers do not disclose their labour income. Then, a number of methods (*e.g.*, deletion of cases with a missing income, OLS and two-step regression imputation, hot-deck) are used to impute the missing incomes. Because the exact value of workers' labour incomes is known (as nonresponse is simulated from full-information samples), it is possible to compute "true" and

[1] Feres (1998), Table 3.

"imputed" inequality coefficients and assess the ability of the different methods to give unbiased inequality estimates. As is demonstrated, the use of inadequate methods to correct for income nonresponse can eventually produce biased inequality coefficients, which magnitude would depend not only on the correction method used but also on the inequality coefficient.

A number of caveats applies. First and as was said above, income nonresponse is not the only measurement error in surveys and may not be the most serious one in terms of its impact on inequality measurement. Income underreporting posses more challenges both in terms of its detection and in terms of its actual impact on income-based statistics. Nonresponse was chosen over underreporting because its correction is more frequent and more methods have been used to deal with it. In the simulations that follow, it is assumed that income non-response is the only measurement error present in the data. Second, it is not the objective of this paper to provide the best possible method to correct nonresponse. Instead, the objective is to assess the biases on inequality measures that methods frequently used to correct for nonresponse may introduce. Third, the discussion that follows should be put into the context of using income (or, expenditure) as an imperfect measure of welfare (often the ultimate variable of interest). Trying to measure incomes without errors does not overcome the issue of incomes being only an imperfect proxy for welfare, a multidimensional and complex concept (on this, see Kakwani and Silber, 2008).

The essay is structured as follows. The next Section presents a discussion of the effects that different patterns of income nonresponse may have on income inequality coefficients. In Section 3 several correction methods commonly found in the empirical literature are described. Section 4 explains the simulations in the context of a particular survey, the Argentinean Permanent Household Survey, while in Section 5 the results of the simulations are presented and discussed. Section 6 presents the conclusions of the essay.

II. MEASURES OF INEQUALITY AND INCOME NONRESPONSE

There is not a single best inequality coefficient to measure inequality for the simple reason that there is not a single dimension or aspect of inequality to be measured. One might be interested, for instance, in measuring the maximum income distance between the poorest and the richest individuals in a society or, alternatively, one might consider as a pertinent inequality dimension the extent of the income dispersion at a specific part of the income distribution (*e.g.* the lower tail). Furthermore, inequality coefficients may not only differ in how they consider descriptive aspects of inequality (*e.g.* the distance between the highest and the lowest income), but also in how they assess subjective or normative dimensions associated with ethical values and attitudes towards inequality itself. Even inequality coefficients that describe objective aspects of inequality, such as the well-known Gini coefficient, have attached ethical dimensions (Cowell, 2000). The existence of a large number of inequality coefficients is thus justified: all of them take a particular aspect of inequality (*e.g.* the income dispersion at the lower tail of the distribution) and summarise it in a number, suitable to make income distribution comparisons possible.

For these reasons and following Champernowne (1974), four different inequality measures are used. The first one is the well-known Gini coefficient which, as demonstrated by Champernowne (1974), is sensitive to inequality around the mean of the income distribution (*i.e.* "inequality among the less extreme incomes"). The second is the Theil index (a member of the family of General Enthropy measures), which unlike the Gini coefficient is relatively more sensitive to income dispersion at the upper tail of the distribution (Champernowne, 1974). Finally, the third and fourth ones are versions of the Atkinson index with inequality-aversion parameters equal to 1 (gives an inequality coefficient that is relatively sensitive to inequality at the top of the income distribution) and 2 (gives a coefficient relatively sensitive not only to dispersion at the lower part of the distribution, but also to the existence of extremely low income values).

2.1. Patterns of Income Nonresponse

Income nonresponse occurs when individuals refuse to answer income questions in a survey. The reasons for such a refusal can be varied, although it is generally acknowledged that a strong motive is the fear of a negative fiscal reaction (*e.g.* the loss of a subsidy, more taxation) to the potential disclosure of income. The statistical relationship between the probability of not responding to income questions and the distribution of certain explanatory variables including the actual income of surveyed individuals is called the pattern of missingness or the pattern of income nonresponse.

Let us arrange the survey data in a matrix $\mathbf{M} = \{m_{ij}\}$, $i = 1, \ldots, n$; $j = 1, \ldots, k$, where i indexes individual data and j the variables collected for each individual. $\mathbf{M}$ can be partitioned into two sub-matrices: one full data matrix, $\mathbf{M}_l$, where all variables for each individual are fully observed, and a matrix $\mathbf{M}_{n-l}$, where income data (one of the columns of $\mathbf{M}$) is missing for each individual. Additionally, let $\mathbf{Q} = \{q_{ij}\}$, denote a n x k matrix, where $q_{ij} = 1$ if m_{ij} is observed and zero otherwise. The joint distribution of $\mathbf{M}$ and $\mathbf{Q}$ takes the general form

$$f(\mathbf{M},\mathbf{Q} \mid \boldsymbol{\theta},\boldsymbol{\beta}) = f(\mathbf{M} \mid \boldsymbol{\theta}) f(\mathbf{Q} \mid \mathbf{M},\boldsymbol{\beta}) \qquad (1)$$

In expression (1), $f(\mathbf{Q} \mid \mathbf{M},\boldsymbol{\beta})$ denotes the distribution of the pattern of missingness, and θ and β are vectors of parameters (Weeks, 2001). The literature on nonresponse recognises that such a pattern can take three forms:

a) Missing completely at random *(mcar):* $\mathbf{M}_l$, the complete-case sub-matrix, is formed by a random sub-sample of $\mathbf{M}$. The pattern of missing data cannot be predicted from any sample information, as it depends neither on the distribution of the missing data nor on the distribution of the observed data. In terms of expression (1), $f(\mathbf{Q} \mid \mathbf{M}_l,\mathbf{M}_{n-l},\boldsymbol{\beta}) = f(\mathbf{Q} \mid \boldsymbol{\beta})$. In other words, the distribution of missing data does not depend on any explanatory variable (*e.g.* gender, age, educational attainments, etc.) nor on the value of the missing variable (income).

A *mcar* pattern corresponds to one where the distribution of missingness is, for example, the same for males and females, for the different educational and age categories, etc. In this situation, such a distribution could not be predicted from the distribution of those variables. As King *et al.* (2001) point out this situation would be equivalent to one in which individuals decide whether to answer income questions or not on the basis of the flip of a coin.

b) Missing at random *(mar):* "if the conditional probability of the observed pattern of missing data, given the missing data and the value of the observed data (*i.e.*, $f(\mathbf{Q} \mid \mathbf{M}_l, \mathbf{M}_{n-l}, \boldsymbol{\beta})$), is the same for all possible values of the missing data such that $f(\mathbf{Q} \mid \mathbf{M}_l, \mathbf{M}_{n-l}, \boldsymbol{\beta}) = f(\mathbf{Q} \mid \mathbf{M}_l, \boldsymbol{\beta})$" (Weeks, 2001). In other words, the distribution of the missing data is conditional only on the distribution of the observed data. Thus, the pattern of missingness can be predicted from the distribution of certain explanatory variables (*e.g.* gender, age, educational attainments, etc.) but is independent from the actual value of the missing variable (income). In this situation, for instance, missingness would be more frequent among males than among females, or would be associated with the age of the surveyed individuals.

c) Non-ignorable *(ni)* nonresponse: occurs whenever the distribution of **Q** depends on the actual value of the variable with missing cases (*e.g.* income). In other words, the probability of income being missing for a particular individual (*e.g.* the $(l + k)^{th}$ individual) depends on her actual income (y_{l+k}).

Correction methods assume a pattern of missingness in order to impute missing incomes. If the pattern assumed by a particular correction method does not coincide with the actual one, incomes may be imputed with error. If this is the case, inequality coefficients will be estimated with error. In general, the magnitude of the error in inequality estimation will depend upon two factors:

1) for a given inequality coefficient (*e.g.* Gini, Theil, Atkinson indexes), the discrepancy between the actual pattern of income missingness and the one assumed by the correction method;
2) for a given correction method, the existing relationship between the aspect of the inequality considered by a particular coefficient (as explained in Section 2) and the actual pattern of missingness.

The first factor refers to an issue which is usually highlighted by the literature: correction methods are relatively effective in producing accurate imputations only when the pattern of missingness they assume coincides with the actual one (*e.g.* Greenlees *et al.*, 1982; Little and Rubin, 1987). If this is not the case, any particular inequality coefficient (*e.g.* Gini, Theil, Atkinson indexes), disregarding the aspect of inequality they measure, will be estimated with error. The second factor, which is related to the specific characteristics of each inequality coefficient, has not been addressed before in the literature (as far as the authors know) and highlights the fact that inequality coefficients are sensitive to income dispersion at specific parts of the distribution (*e.g.* at the upper tail). If incomes are missing, for instance, at the upper tail, any correction method will have a larger impact on coefficients that are sensitive to income dispersion at that specific part of the distribution (*e.g.* the Theil coefficient). Thus, imputation errors introduced by correction methods will be reflected more importantly on coefficients that are sensitive to dispersion at the part of the distribution where missing incomes are being imputed.

Both factors are equally important and, when ignored, contribute to the distortion of the estimation and interpretation of inequality coefficients. In the next Section, a number of correction methods are described. Rather than listing all the methods proposed in the theoretical literature on missingness, a core of methods commonly found in the empirical literature on inequality is presented.

III. METHODS TO CORRECT FOR INCOME NONRESPONSE

There are basically two alternative ways that can be used to handle the income nonresponse problem: (i) the removal of cases with missing incomes from the sample; and (ii) the imputation of an income to them. The first alternative, the deletion of cases with missing income, constitutes the simplest way of handling this problem and, as such, has been widely used in empirical papers analysing income distribution in countries with high nonresponse rates (Altimir and Beccaria, 1999; Gasparini *et al.,* 2001, etc.). It consists in removing from the original sample all the cases where incomes are missing and analysing the remaining complete-case sub-sample as if it were representative of the population.[2] The key assumption in this methodology is that removed cases are a random sub-sample of the whole sample or, in other words, that the pattern of missingness is *mcar* (as defined in Section 2.1). If, in fact, this is the case, the deletion of observations with missing income does not introduce any bias in the estimation of inequality coefficients (as is shown in Section 5). Nevertheless, even in such a case this method implies the loss of valuable information related to other variables (*e.g.* age, educational attainment, type of job) that are discarded when individuals do not respond to income questions. In addition, the reduction of cases to only complete ones produces an overestimation of the confidence with which inferences are undertaken (King *et al.*, 2001). The performance of deletion is worse if the actual pattern of missingness is not *mcar.* With either *mar* or, especially, *non-ignorable* nonresponse the deletion of cases with missing incomes produces biased estimates of income statistics (*e.g.* mean, variance; Little and Rubin, 1987; King *et al.,* 2001) and of specific inequality coefficients (see Section 5).

For the second alternative, the imputation of missing incomes, a number of different methods have been proposed. Some of them are parametric (*e.g.* OLS models and two-step regressions, maximum likelihood methods), assuming an underlying population function between income (the variable to be imputed) and a number of individual characteristics (such as age, gender, educational attainment, *etc.*), while

[2] There are two types of deletion of cases with missing income. The first one is the simple removal of such cases from the dataset. The second one is the removal of cases with missing incomes and the reweighing of the remaining complete-case sub-sample to keep it representative of the population (see Little and Rubin, 1987).

others are non-parametric (*e.g.* cold-deck, hot-deck). All of them assume a particular pattern of missingness in order to make imputations. In most empirical studies, assumptions on the pattern of missingness are untestable as there exists only one source of information for incomes (*e.g.* household surveys), so when they are missing there is no way of knowing what is the statistical relationship between the probability of them being missing and the explanatory variables (including income itself). In the literature reviewed only a few papers (Greenlees *et al.*, 1982, for instance) can exactly identify the actual pattern of missingness (as the authors have two datasets formed by the same households, one where all the variables are fully observed, and the other one with missing income cases).

The following are the main imputation methods found in this empirical literature:

a) *Standard OLS models*: this method assumes the existence of an underlying linear population relationship between income (the variable with missing data) and a set of explanatory variables (*e.g.* gender, age, education, marital status) fully observed from individuals. After estimating such a relationship for the complete-case sub-sample, the missing incomes are imputed using the estimated parameters from the complete-case sub-sample.[3] This method assumes that the pattern of missingness is *mar* or, in other words, that the distribution of missing incomes is conditional on the distribution of data observed for the set of explanatory variables and can be predicted from it. When this is the case, standard OLS models produce unbiased estimates of the mean income though estimates of the variance are downward biased. The reason is that imputations come from a regression to the mean and do not reflect the actual variation in the distribution of **y** given $\mathbf{M}_{-1}$. This makes standard OLS models inappropriate for imputing missing incomes for the study of inequality, as this precisely requires accurate estimates of income dispersion. When the actual pattern of missingness is *non-ignorable* this method produces biased estimates of income mean and variance (Greenlees *et al.*, 1982). Inequality coefficients may also be estimated with large errors (see Section 5). Despite its limitations, standard OLS regression models have been widely used in the empirical literature. For instance, Székely and Hilgert (2007) for 18 Latin American countries, Gasparini (1999) and Gasparini and Sosa (1999) in the Argentinean case, Larrañaga (1999) for Chile, among others, use standard OLS models to impute missing incomes. In all these cases, the authors do not have any additional information to infer the actual pattern of missingness, assuming that it is *mar*.

[3] Let us partition matrix **M** (as defined in Section) into two matrices: $\mathbf{M}_{-1}$, a n x $(k-1)$ matrix containing full information for the variables that form it (all variables but income); and **y** a n x 1 vector containing information on income. Only a submatrix of **y** (the first l cases) is complete, while the remaining ($n-l$ cases) are missing. Standard OLS models assume that the underlying population function is given by $\mathbf{y}=\mathbf{M}_{-1}\boldsymbol{\theta}+\varepsilon$, where ε is $iid(0,\sigma^2)$. Thus, this method consists in using the expectation $E_l(\mathbf{y}_{n-l} \mid \mathbf{M}_{-1})=\mathbf{M}_{-1,n-l}\boldsymbol{\theta}$ to estimate $\widehat{E_l}(\mathbf{y}_{n-l} \mid \mathbf{M}_{-1})=\mathbf{M}_{-1,n-l}\hat{\theta}$, where $\widehat{E_l}(.)$, where indicates that the expectation is taken using information available for the complete l cases (Weeks, 2001).

b) *Random OLS models*: this method is similar to standard OLS models in its assumption of the underlying population relationship between income and a set of explanatory variables (*e.g.* gender, age, education, marital status) fully observed from individuals. Unlike standard OLS models, random OLS models produce better estimates of the variance of the imputed variable (income) by adding a stochastic error to the equation imputing incomes. Thus this method estimates a regression between income and a set of explanatory variables for the complete-case sub-sample (as standard OLS models do). Missing incomes are imputed using the estimated parameters in the complete-case regression plus a random error term that can be obtained from a normal distribution with a zero mean and a standard deviation equal to the standard deviation of the complete-case regression. Thus, the $(l + 1)^{th}$ imputed value is

$$\hat{y}_{l+1} = \mathbf{M}_{-1,l+1}\hat{\boldsymbol{\theta}} + \hat{e}_{l+1}$$

where $\hat{e}_{l+1} \sim \mathbf{N}\left(0, \hat{\boldsymbol{\sigma}}_l^2\right)$ and $\hat{\boldsymbol{\sigma}}_l^2 = \sum_{i=1}^{1} \frac{\left(y_i - \hat{y}_i\right)}{l}$ (*e.g.* standard error of the regression estimated on the complete cases). Another option might be to randomly sample (with replacement) the observed residuals from the complete-case regression and use them as the random error term (Weeks, 2001). Like standard OLS models, this method assumes that the pattern of missingness is *mar* (Little and Rubin, 1990).

c) *Hot-deck*: there are several variants for imputing missing values using this non-parametric method. The simplest one consists in partitioning the data into non-overlapping “cells” according to determined characteristics (*e.g.* gender, age, educational level, working sector) and allocating individuals (respondents and non-respondents) to these cells. After this allocation is made, respondent individuals within each cell are chosen randomly and with replacement to “donate” their incomes to non-respondents in the same cell. This process can be applied once or several times to produce a set of values for each non-respondent. An appealing characteristic of this method is that it does not presuppose the existence of any underlying population function to impute missing incomes. The pattern of missingness assumed is *mar*, as it is presupposed that the probability of nonresponse may vary across cells but not within them. When the actual pattern of missingness is *mar* and the number of donors (*i.e.* respondent individuals) within each cell is large with respect to nonrespondents, the hot-deck gives unbiased estimates of mean income and variance (Little and Rubin, 1990). Alternatively, if the actual pattern of missingness is *non-ignorable*, the hot-deck produces biased estimations of mean and variance (Greenlees *et al.*, 1982). Different variants of this method are used, for instance, to impute missing incomes in the US Current Population Survey and by most OECD Statistical Offices (Atkinson *et al.*, 1995). Ruiz-Tagle (1998) uses this method to impute missing incomes in the Chilean household survey, Biewen

(1999) uses it with German household data and Banks *et al.* (2002) use a hot-deck to impute financial wealth in Great Britain. Some National Statistical Offices use variations of the hot-deck. For instance, the Chilean Statistical Office uses the cold-deck, which instead of randomly sampling (with replacement) individuals to donate their incomes to nonrespondents, imputes the mean value of respondents' incomes within each cell to nonrespondents (Ruiz-Tagle, 1998). Lillard *et al.* (1986) criticises its use in the US Current Population Survey, finding it produces biased estimations of mean incomes, whereas Paulin and Ferraro (1994) conclude that it is not possible to use hot-deck in the US Consumer Expenditure Survey as the samples of this survey are too small for its effective use.

d) *Two-step regression models*: these models assume that incomes are reported only when the utility for individuals of answering income questions is positive. It is assumed that such an utility level depends on income, determining that the probability of answering income questions depends on income itself. In the first step of these models, a "reporting-decision equation", where a dependent dichotomous variable (*i.e.* to report income or not) is regressed on a set of explanatory variables, is estimated. In the second step, the decision made as to whether to answer or not the income questions is considered to impute incomes to nonrespondents. The models are as follows. Incomes are observed only if $\mathbf{X_2}\boldsymbol{\beta}_2 + u_2 > 0$, with a probability of $\Phi(\mathbf{X_2}\boldsymbol{\beta}_2)$, where $u_2 \sim N(0,1)$. There exists an underlying population function, relating incomes with a set of explanatory variables, of the form:

$$y = \mathbf{X_1}\boldsymbol{\beta}_1 + u_1$$

where $u_1 \sim N(0,\sigma^2)$. It is also assumed that (u_1, u_2) ~bivariate normal $(0,0,\sigma^2,1,\rho)$. Thus,

$$E\left(y \mid \mathbf{X_2}\boldsymbol{\beta}_2 + u_2 > 0\right) = \mathbf{X_1}\boldsymbol{\beta}_1 + \rho\sigma^2\lambda(\mathbf{X_2}\boldsymbol{\beta}_2)$$

where $\lambda(\mathbf{X_2}\boldsymbol{\beta}_2) = \frac{\phi(\mathbf{X_2}\boldsymbol{\beta}_2)}{\Phi(\mathbf{X_2}\boldsymbol{\beta}_2)}$.

As it is discussed in Wooldridge (2002), bivariate normality of the errors may be an overly restrictive condition and indeed, a normality assumption on u_1 is not even needed. Two-step regression models are relatively frequent in selection-bias contexts (Heckman, 1979) and, in this case, can impute incomes without biases with a *non-ignorable* pattern of missingness. However, special care has to be put into their econometric specification, since it has been noticed that they may be highly sensitive to model misspecification particularly regarding the division of the total set of covariates into those being used in the first step (*e.g.* the "reporting-decision equation") and the second step (Weeks, 2001). Paulin and Ferraro (1994), for instance, criticise Greenlees *et al.* (1982) findings of a *non-ignorable* pattern of missingness in a US household survey and attribute such findings to an underspecification of the model. With a suitable model specification, they obtain a *mar* pattern of missingness and find that imputations using the two-step regression model produce biased mean income estimates. A number of empirical papers use these models to impute missing data. For instance, Biewen (1999) uses one to correct German data and Greenlees *et al.* (1982) and Lillard *et al.* (1986) employ different two-step regression models on US data.

Other correction methods have been proposed but are rarely found in the empirical literature due to the specific types of missingness they address. An example is Victoria-Feser (2000), which analyses the robustness of estimates of parameters of an income distribution (*e.g.* parameters of a Gamma) when income data is contaminated due to the truncation of the distribution. In a simulation framework, she considers two imputation methods (a classical maximum likelihood method and a modification of M-estimator methods) and compares the performance of both methods, finding that the M-estimator model performs better.

IV. A SIMULATION EXERCISE

Most of the empirical papers which estimate inequality from household survey data correct for income nonresponse without considering how the choice of an inadequate correction method affects the inequality measurement. The lack of administrative or tax data (or any other source of information other than household surveys) hampers any attempt to compare the methods' performances in producing accurate inequality estimates. Furthermore, in most of the cases, it is not even possible to be sure whether the inequality estimates are under or overestimated. Given the lack of information on both the effects of particular correction methods on inequality coefficients and on the pattern of missingness followed by the data, a valid option is to estimate *bounds* for the errors in the estimation of such coefficients when the data follows certain patterns of missingness. This can be done by using a simulation framework.[4] In such simulations, an income distribution can be artificially contaminated with a pattern of missingness and then corrected by any of the methods described in Section 3. After doing that, any inequality coefficient (for instance, those described in Section 2) can be estimated over the corrected distribution and compared with the ones calculated over the actual distribution. In this way, the effects of using different correction methods on the inequality measurement can be examined. This exercise is carried out here.

One option in performing this exercise is to simulate a general or theoretical income distribution using a standard density function (*e.g.*, Gamma, Pareto, Singh-Maddala, *etc.*), to contaminate it and to correct it using several correction methods. However,

[4] Simulations or counter-factual exercises are frequent in inequality analysis with contaminated data. To the aforementioned paper of Victoria-Feser (2000), one can add Cowell and Victoria-Feser (2001) where the authors artificially trim income distributions (at the top or at the bottom of the distribution) to test the robustness of partial welfare orderings to the presence of contaminated data at the extremes of the distributions. Other papers using income distribution simulations are Champernowne (1974), Cowell and Victoria-Feser (1996), Cowell *et al.* (1999) and Cowell and Flachaire (2002).

in the context of this exercise, one would have to simulate not only the income density function but also the density functions of all the variables related to income (*e.g.* age, gender, educational level) that are used by different correction methods (*e.g.* OLS models, hot-deck) to impute missing incomes. A clear drawback of this option is that justifying all the assumptions required to construct the simulation exercise may interfere with the simple purpose of assessing the performance of the correction methods.

The other option would be to work with actual data on income and other variables (*e.g.* age, gender, educational level), gathered by a specific household survey. On such data, several patterns of missingness (*e.g. mcar*, *mar*, *non-ignorable*) can be simulated. After doing this, missing incomes can be imputed and inequality coefficients estimated using both true and corrected incomes. The main advantage of this option is that it allows us to concentrate directly on the relationship among patterns of missingness, correction methods and inequality measurement, without worrying about modelling issues. This is achieved, however, at the cost of making the analysis and the results specific to the type of data used as inputs for these exercises. This cost can be made less onerous if several different data sets (*i.e.* for the same country but for different years or for different countries) are used.

In this essay, we follow this last option by using data from the Argentinian Permanent Household Survey (henceforth, PHS). This choice is made on several grounds. First, the PHS is the only source of information in Argentina which can be used to measure income inequality. There are no other sources (*e.g.* administrative data, tax records) which can be used to infer missing incomes, to correct for data problems, or to speculate about the actual pattern of missingness existing in the PHS. Second, the PHS is one of the surveys with the highest income nonresponse rates in Latin America, a region with particularly low standards in statistical information (Feres, 1998; Széekely and Hilgert, 2007). In this respect, to work on the PHS is to consider a survey where, given the magnitude of the nonresponse rate, the missing income problem cannot be ignored in the course of studying inequality.

The data used in this exercise are samples of wage earners, who report their labour incomes and other relevant variables (*e.g.* age, gender, educational level, etc). Wage earners have been selected (instead of entrepreneurs or the self-employed, for instance) as they constitute the largest labour category (approximately 3000 observations per survey) and the one less subject to income fluctuations. Two years, 1995 and 1998, have been chosen. Both these years were years of relatively high inequality in reported incomes and both were years of average income nonresponse. Three patterns of missingness have been simulated:

a) a *mcar* process, where all the individuals in the sample have the same probability of being chosen as nonrespondent individuals. This process is simulated by drawing random numbers from a uniform distribution and assigning them to each individual in the sample. Then, if the intention is to consider, for instance, a situation where 12% of the sample have missing incomes, individuals with random numbers higher or equal to 0.88 have their incomes deleted.

b) a *non-ignorable* nonresponse process, where only individuals located at the lower tail of the actual income distribution face the same probability of being chosen as nonrespondent individuals. To generate cases with missing incomes, random numbers drawn from a uniform distribution are assigned to each individual at the lower tail of the distribution. If, for instance, it is simulated that 12% of the cases in the sample have missing incomes, random numbers are drawn for the lowest 24% incomes and then half of them are randomly deleted. This procedure is adopted to avoid truncating the income distribution.

c) a *non-ignorable* nonresponse process, where only individuals located at the upper tail of the actual income distribution face the same probability of being chosen as non-reporting individuals. The process of generating cases with missing incomes is similar to the one described in b).

Regarding the magnitude of income nonresponse, two scenarios have been considered. In the first one, it is assumed that 7% of the population does not answer income questions. The second one assumes that such a share is 12%. The actual proportion of wage earners not responding to income questions in the PHS during the last decade has oscillated between 7% and 12%. Thus, the first scenario can be regarded as a low income nonresponse scenario, in the context of the Argentinian PHS, while the second can be considered as a high income nonresponse scenario. All simulations (*i.e.* patterns of missingness) in both scenarios have been replicated 100 times. In each of the 100 rounds of each of the simulations, the data is contaminated (deleting incomes of certain individuals, as described) and the different correction methods are applied. This means that 2400 sets of imputations have been done for each year (100 sets for 4 correction methods, under 3 patterns of missingness for 2 scenarios). For multiple imputation methods (*e.g.* hot-deck) this implies producing 500 complete datasets (if, as is done here, 5 datasets per simulation are produced) per year/simulation/scenario.

An important caveat of this simulation exercise is related to the decision of assuming that missingness occurs at both ends of the income distribution. It has to be recognised that nonresponse patterns existing in actual data are probably more complex than these simulated here. However, these simulations illustrate the effects that missingness can have on inequality measurement (under two extreme forms of missingness). In addition, they show how the choice of a particular correction method (and the inequality coefficient) may affect such a measurement.

4.1. The Correction Methods

Five correction methods, described in Section 3, are used in the simulations. The first one is the deletion of cases with missing incomes. No re-weighing of the remaining complete-case sub-sample is done after the deletion and inequality coefficients are calculated using such a sub-sample (containing only individuals with full information on labour incomes).

The second correction method used is a standard OLS regression model. First, the following standard OLS regression is estimated from the complete-case sub-sample:

$$\begin{aligned} w_i = {} & \alpha_0 + \alpha_1 age_i + \alpha_2 age_i^2 + \alpha_3 marital_i + \alpha_4 sex_i + \alpha_5 educ1_i \\ & + \alpha_6 educ2_i + \alpha_7 skill_i + \alpha_8 tenure_i + \alpha_9 size_i + \alpha_{10} socben_i \\ & + \sum_{w=11}^{20} \alpha_w k_{wi} + e_i \end{aligned} \tag{2}$$

where w_i is the log of the hourly wages and the independent variables are age, squared age, marital status, gender, two dummies for educational attainments (*educ*1 equal to 1 if the individual has up to secondary incomplete, and *educ*2 equal to 1 if she has up to tertiary education incomplete), skill category (non-skilled and skilled), tenure at work, size of the firm, a dummy for receipt of social benefits and a set of ten dummies for different economic sectors (k_w), respectively. Missing incomes are imputed using the $\hat{\alpha}_i$ parameters estimated running regression (2) to impute labour incomes for individuals who are simulated as not reporting them:

$$\begin{aligned} w_j = {} & \hat{\alpha}_0 + \hat{\alpha}_1 age_j + \hat{\alpha}_2 age_j^2 + \hat{\alpha}_3 marital_j + \hat{\alpha}_4 sex_j + \hat{\alpha}_5 educ1_j \\ & + \hat{\alpha}_6 educ2_j + \hat{\alpha}_7 skill_j + \hat{\alpha}_8 tenure_j + \hat{\alpha}_9 size_j + \hat{\alpha}_{10} socben_j \\ & + \sum_{w=11}^{20} \hat{\alpha}_w k_{wj} \end{aligned} \tag{3}$$

The third correction method used is a random OLS model. As described in Section 3, this method is similar to a standard OLS model but it adds a stochastic error in the imputation step. Thus, this method estimates equation (2) but imputes missing labour incomes using:

$$\begin{aligned} w_j = {} & \hat{\alpha}_0 + \hat{\alpha}_1 age_j + \hat{\alpha}_2 age_j^2 + \hat{\alpha}_3 marital_j + \hat{\alpha}_4 sex_j + \hat{\alpha}_5 educ1_j \\ & + \hat{\alpha}_6 educ2_j + \hat{\alpha}_7 skill_j + \hat{\alpha}_8 tenure_j + \hat{\alpha}_9 size_j + \hat{\alpha}_{10} socben_j \\ & + \sum_{w=11}^{20} \hat{\alpha}_w k_{wj} + \hat{e}_j \end{aligned} \tag{4}$$

where variables are as in equation (2) and $\hat{e}_j$ is a stochastic term obtained as a random draw from a $N\left(0, \hat{\sigma}_l^2\right)$, where $\hat{\sigma}_l^2$ is the variance of the regression error (estimated from equation (2)).

The fourth correction method is the hot-deck. The variables used to construct the cells to resample from are gender, marital status (single or not), age (four categories: 15-25, 25-40, 40-65 and over 65) and education (incomplete secondary education, incomplete university education and completed university education). It is not possible in this exercise to use a larger number of categories because of the reduced sample size. Having more categories would imply having cells with an insufficient number of individuals to act as "donors" for missing cases. To partially compensate for this limitation, a multiple-imputation hot-deck is adopted. Thus, once cells are constructed, sets of five imputed incomes are assigned to each nonrespondent individual. As King *et al.* (2001) explain, "multiple imputation involves imputing m values for each missing item and creating m completed data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with different imputations to reflect uncertainty levels." It can be demonstrated that m can be as low as 5 to obtain efficient estimators (*ibid.*, p. 56). Each of these five imputations are incomes donated by reporting individuals (with replacement) within each cell.

Finally, the fifth correction method considered to impute missing incomes is a two-step regression model (à-la-Heckman). The first step estimates a "reporting-decision equation" over the entire sample of individuals:

$$\begin{aligned} z_i &= \delta_0 + \delta_1 age_i + \delta_2 sex_i + \delta_3 marital_i + \delta_4 educ1_i + \delta_5 educ2_i \\ &\quad + \delta_6 hours_i + u_{1i} \end{aligned}$$

where z is a dichotomous variable (0 for individuals not reporting income; 1 for individuals reporting income) and the explanatory variables that influence such a decision are individuals' age, gender, marital status, educational attainments (*educ*1 and *educ*2, as defined for equation (2)) and the number of hours worked per month (*hours*). The second step estimates a wage equation corrected by the decision of not reporting to impute labour incomes for those not reporting it:

$$\begin{aligned} w_j &= \beta_0 + \beta_1 age_j + \beta_2 age_j^2 + \beta_3 sex_j + \beta_4 marital_j + \beta_5 educ1_j \\ &\quad + \beta_6 educ2_j + \beta_7 skill_j + \beta_8 tenure_j + \beta_9 size_j + \beta_{10} socben_j + u_{2j} \end{aligned} \tag{5}$$

where variables in equation (5) are defined as in equation (2).[5] As there are a relatively large number of different variables in the two equations, multicollinearity in the second stage are avoided. Moreover, by including the number of hours worked in a month in the first stage but not in the second, issues of identification are controlled. Theoretically, the number of hours worked in a month may affect the decision of reporting income (for instance, individuals may want to conceal their incomes if they are high, which is conditionally affected by the number of hours worked in a month) without affecting

[5] Considering both steps, equation 5 is

$$\begin{aligned} w_j &= \beta_0 + \beta_1 age_j + \beta_2 age_j^2 + \beta_3 sex_j + \beta_4 marital_j + \beta_5 educ1_j \\ &\quad + \beta_6 educ2_j + \beta_7 skill_j + \beta_8 tenure_j + \beta_9 size_j + \beta_{10} socben_j + \gamma\lambda z_i \end{aligned}$$

where $\gamma = \frac{E(u_2 \mid u_1)}{u_1}$ and λ is the inverse Mill's ratio.

the hourly wage. The parameter showing selection bias (λ in Section 3) is strongly significant in all the estimated cases.

After missing labour incomes are imputed (or individuals with missing incomes are deleted from the sample) four inequality coefficients are calculated for each simulation: the Gini coefficient, the Theil index, and the Atkinson index with $e = 1,2$. Additionally, another measure that considers errors in the imputation process is estimated in each case. It is the Mean Absolute Percentage Error (henceforth MAPE) and is defined as:

$$MAPE = \sum_{i=l+1}^{n} \frac{\frac{\left|y_{i,true} - y_{i,imp}\right|}{y_{i,true}}}{n-l} \times 100 \qquad (6)$$

where $n - l$ is the number of cases with missing income, $y_{i,true}$ is the true income of individual i whose income is simulated as missing and $y_{i,imp}$ is the imputed income obtained from the correction method used. This measure would be zero when a method imputes exactly the actual income values and would grow as imputed incomes differ from the actual ones. It should be noticed that because imputation errors are expressed in percentage terms it is likely that MAPE are higher when incomes are missing at the lower tail of the distribution. Thus, MAPE should be compared across correction methods within each simulated pattern of missingness (*e.g.* incomes missing at the lower tail of the distribution) but not across them.

V. THE RESULTS

Tables 1 and 2 present the ratio between inequality coefficients obtained after imputing using the correction methods presented in Section 4.1 and the actual inequality coefficients, when income data is *mcar*. The two-step regression is not considered, as this method reduces to a standard OLS model when income is *mcar* (by definition - see Section 2.1 - under *mcar* no variable can explain the decision of not reporting labour income). Table 1 shows the first scenario (proportion of income missingness equal to 7% of the total population), while Table 2 presents the results for the second scenario (proportion of income missingness equal to 12% of the total population).

Both tables show that when income is *mcar* most correction methods provide accurate estimations of the several inequality coefficients considered. Statistically, only inequality coefficients obtained after applying the standard OLS model are significantly lower than the actual ones, whereas inequality coefficients obtained after using the random OLS model are significantly higher than the actual ones (in all cases except for the Theil coefficient). As explained in Section 3, the standard OLS model produces downward-biased estimations of the standard deviation of the variable being imputed, which in turn produces biased inequality estimations. A lower income standard deviation has an effect on all the inequality coefficients, though some of them are more affected than others. For instance, the Theil coefficient (Part

TABLE 1

EFFECTS OF NONRESPONSE UNDER *MCAR* WHEN THE PROPORTION OF MISSINGNESS IS 7%

A. Gini coefficient		
Method	1995	1998
Deleting	0.9996	1.0002
Simple regression	0.9863	0.9874
Random regression	1.0055	1.0039
Hotdeck	0.9997	1.0003
True distribution	0.3881	0.4103

B. Theil coefficient		
Method	1995	1998
Deleting	0.9992	1.0008
Simple regression	0.9697	0.9732
Random regression	1.0007	1.0002
Hotdeck	0.9996	1.0009
True distribution	0.2931	0.3119

C. Atkinson index (*e* = 1)		
Method	1995	1998
Deleting	0.9994	1.0004
Simple regression	0.9798	0.9828
Random regression	1.0117	1.0100
Hotdeck	0.9998	1.0005
True distribution	0.2329	0.2612

D. Atkinson index (*e* = 2)		
Method	1995	1998
Deleting	0.9992	1.0001
Simple regression	0.9912	0.9958
Random regression	1.0184	1.0171
Hotdeck	1.0001	1.0002
True distribution	0.4229	0.4872

Ratios between inequality measure after imputing and the actual one (true distribution). A value higher than one implies an overestimation of the true inequality level, whereas a value lower than one implies an underestimation of this level.
Figures in italics are statistically non-significant (95%).

TABLE 2

EFFECTS OF NONRESPONSE UNDER *MCAR* WHEN THE PROPORTION OF MISSINGNESS IS 12%

A. Gini coefficient		
Method	1995	1998
Deleting	1.0001	0.9996
Simple regression	0.9772	0.9779
Random regression	1.0100	1.0072
Hotdeck	1.0008	0.9994
True distribution	0.3881	0.4103

B. Theil coefficient		
Method	1995	1998
Deleting	0.9998	0.9992
Simple regression	0.9495	0.9524
Random regression	1.0020	1.0023
Hotdeck	1.0023	0.9988
True distribution	0.2931	0.3119

C. Atkinson index (*e* = 1)		
Method	1995	1998
Deleting	1.0002	0.9993
Simple regression	0.9670	0.9694
Random regression	1.0205	1.0172
Hotdeck	1.0017	0.9991
True distribution	0.2329	0.2612

D. Atkinson index (*e* = 2)		
Method	1995	1998
Deleting	1.0004	0.9993
Simple regression	0.9878	0.9905
Random regression	1.0310	1.0275
Hotdeck	1.0015	0.9992
True distribution	0.4229	0.4872

Ratios between inequality measure after imputing and the actual one (true distribution). A value higher than one implies an overestimation of the true inequality level, whereas a value lower than one implies an underestimation of this level.
Figures in italics are statistically non-significant (95%).

B of Tables 1 and 2) and the Atkinson ($e = 1$) index (Part C of Tables 1 and 2) are more affected than the Gini coefficient and the Atkinson ($e = 2$) index (Parts A and D of Tables 1 and 2, respectively), as income is relatively more dispersed at the upper tail of the income distribution, which means that income dispersion in that part of the distribution, after imputing with the standard OLS model, will suffer the largest underestimation. Accordingly, inequality coefficients which are sensitive to income dispersion in that part of the distribution, such as the Theil and the Atkinson ($e = 1$) indexes will be the most affected. Using the random OLS imputation method corrects that underestimation.

From the perspective of a policy-maker or an analyst, all these methods produce relatively good estimations. Even standard OLS estimates for the Gini are less than 3% lower than the actual ones (see Table 2, Part A) and for the Theil coefficient are around 5% lower than the actual ones (Table 2, Part B). In the case of the Atkinson indexes, standard OLS produces estimates which are 3% and 1% lower than the actual ones, when $e = 1,2$, respectively (Table 2, Parts C and D).

Another perspective on the performance of the correction methods is obtained by looking at their MAPEs (as defined in expression (6)). Part A of Table 3 shows the MAPEs for the first scenario (7% of missingness), while Part B shows the MAPEs for the second one (12% of missingness). Because income data is randomly missing (from all parts of the distribution) there is no difference across scenarios in the imputation errors introduced by the methods and, consequently, MAPEs for every method are similar across scenarios. Both scenarios show that deleting produces the highest MAPE (around 95%-120%) and that the standard OLS model produces the lowest ones (as it minimises the squared imputation errors), of around 47%. Because

TABLE 3

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE) WHEN DATA IS *MCAR*

a) Proportion of income missingness: 7%

Method	1995	1998
Deleting	94.64%	116.83%
Simple regression	46.71%	46.60%
Random regression	72.66%	73.82%
Hotdeck	86.05%	95.20%

b) Proportion of income missingness: 12%

Method	1995	1998
Deleting	94.09%	117.64%
Simple regression	46.02%	46.94%
Random regression	73.70%	74.60%
Hotdeck	85.49%	96.82%

the random OLS model incorporates a stochastic term ($\hat{e}_j$ in equation (4)), its MAPE is higher than the one for the standard OLS model, though still considerably lower than the one for deletion.

5.1. Incomes Missing at the Lower tail of the Distribution

Tables 4 and 5 show the ratio between inequality coefficients after imputing missing incomes and the actual inequality coefficients when incomes are missing only at the lower tail of the income distribution. Unlike the *mcar* case, inequality coefficients experience large variations across methods and scenarios (*e.g.* 7% and 12% of income missingness). In general, they are underestimated for all correction methods (described in Section 4.1), the two-step regression model being the only exception, which for certain inequality coefficients (*e.g.* the Gini) produces very accurate estimations, while for others (*e.g.* the Atkinson indexes) overestimates true inequality. The random OLS model also produces an overestimated Atkinson (*e* = 2), though the magnitude of the overestimation is smaller than in the two-step regression case.

TABLE 4

EFFECTS OF NONRESPONSE WHEN INCOME IS MISSING AT THE LOWER TAIL AND THE PROPORTION OF MISSINGNESS IS 7% (*)

A. Gini coefficient			B. Theil coefficient		
Method	1995	1998	Method	1995	1998
Deleting	0.9519	0.9475	Deleting	0.9209	0.9082
Simple regression	0.9776	0.9811	Simple regression	0.9655	0.9701
Random regression	0.9822	0.9822	Random regression	0.9701	0.9707
Hotdeck	0.9465	0.9459	Hotdeck	0.9123	0.9082
Two-step regression	0.9981	0.9977	Two-step regression	1.0025	1.0011
True distribution	0.3881	0.4103	True distribution	0.2931	0.3119

C. Atkinson index (*e* = 1)			D. Atkinson index (*e* = 2)		
Method	1995	1998	Method	1995	1998
Deleting	0.8944	0.8819	Deleting	0.8696	0.8657
Simple regression	0.9690	0.9720	Simple regression	0.9914	0.9965
Random regression	0.9832	0.9821	Random regression	1.0279	1.0322
Hotdeck	0.8882	0.8818	Hotdeck	0.8681	0.8687
Two-step regression	1.0355	1.0300	Two-step regression	1.1253	1.1118
True distribution	0.2329	0.2612	True distribution	0.4229	0.4872

Ratios between inequality measure after imputing and the actual one (true distribution). A value higher than one implies an overestimation of the true inequality level, whereas a value lower than one implies an underestimation of this level.
(*) 7% of missingness implies that the missing incomes are located at the lowest 14% of the distribution.
Figures in italics are statistically non-significant (95%).

TABLE 5

EFFECTS OF NONRESPONSE WHEN INCOME IS MISSING AT THE LOWER TAIL AND THE PROPORTION OF MISSINGNESS IS 12% (*)

A. Gini coefficient			B. Theil coefficient		
Method	1995	1998	Method	1995	1998
Deleting	0.9428	0.9336	Deleting	0.9051	0.8842
Simple regression	0.9644	0.9674	Simple regression	0.9437	0.9476
Random regression	0.9762	0.9735	Random regression	0.9558	0.9534
Hotdeck	0.9338	0.9307	Hotdeck	0.8898	0.8832
Two-step regression	1.0144	1.0102	Two-step regression	1.0288	1.0241
True distribution	0.3881	0.4103	True distribution	0.2931	0.3119

C. Atkinson index (*e* = 1)			D. Atkinson index (*e* = 2)		
Method	1995	1998	Method	1995	1998
Deleting	0.8880	0.8691	Deleting	0.8774	0.8718
Simple regression	0.9541	0.9581	Simple regression	0.9853	0.9965
Random regression	0.9837	0.9792	Random regression	1.0431	1.0426
Hotdeck	0.8783	0.8704	Hotdeck	0.8768	0.8802
Two-step regression	1.0841	1.0786	Two-step regression	1.1894	1.1833
True distribution	0.2329	0.2612	True distribution	0.4229	0.4872

Ratios between inequality measure after imputing and the actual one (true distribution). A value higher than one implies an overestimation of the true inequality level, whereas a value lower than one implies an underestimation of this level.
(*) 12% of missingness implies that the missing incomes are located at the lowest 24% of the distribution.
Figures in italics are statistically non-significant (95%).

Deleting cases with missing incomes is a bad choice when such a missingness occurs at the lower tail of the distribution. For instance, deletion produces Gini coefficients that are almost 7% below the true ones (Table 5, part A). In the case of the Theil coefficient (as explained in Section 2, a measure sensitive to dispersion at the top of the distribution), such underestimation can be almost 12% (Table 5, part B). Regarding the Atkinson (*e* = 1,2) indexes, deleting cases with missing income produces an underestimation of up to 13%. Similar results are obtained when the hot-deck is used.

In general, parametric imputation methods, such as the standard and the random OLS models and the two-step regression, produce relatively more accurate estimates of inequality coefficients. Imputing missing incomes using a standard OLS model produces Gini coefficients that are inferior to the original ones, though the underestimation is lower than the one produced by deletion or the hot-deck: when the proportion of missingness is 7%, such underestimation is around 2% and increases to 3% when the proportion of missingness rises to 12%. In the case of the Theil coefficient, the underestimation is around 5%, when the proportion of missingness is 12% (Table 5, part B). While the Atkinson (*e* = 1) index displays similar results to the Theil coefficient, the Atkinson (*e* = 2) index shows almost no difference from the actual one. The random OLS model slightly improves estimations of the Gini, the Theil and the Atkinson (*e* = 1) index. In the case of the two-step regression imputations, the Gini coefficient is accurately estimated, while the Theil coefficient shows a maximum overestimation of 3% (Table 5, Part B). Atkinson indexes are overestimated in all cases and such overestimation can be as high as 18% (Table 5, Part D).

Two important conclusions can be extracted from these results. The first one, related to each method's characteristics, is that parametric methods, such as standard and random OLS models or two-step regression models, generally produce more accurate inequality coefficients than hot-deck and deletion, when incomes are missing at the lower tail of the distribution. Such methods have a higher probability of imputing incomes lower than the actual ones than, for instance, the hot-deck. On average, the hot-deck imputes lower incomes in only 5% of the cases, whereas that figure is close to 30% in the case of the standard OLS model and close to 45% in the case of the two-step regression model. To impute an income which is lower than the actual one means that any of the four inequality coefficients considered here will show a rise in inequality (of different magnitude, according to the coefficient). Thus, methods that impute a larger proportion of incomes lower than the actual ones will result in higher inequality coefficients. While for some coefficients that are relatively sensitive to dispersion at the lower part of the distribution, such as the Atkinson (*e* = 2), the imputation of lower incomes produce strong increases in inequality, eventually leading to an overestimation of it (see Part D in Tables 4 and 5), for others, which are relatively less sensitive to lower tail dispersion (*e.g.* the Gini coefficient), this effect is less important (see part A in Tables 4 and 5).

The second conclusion, related to each inequality coefficient's characteristics, is that coefficients that are relatively sensitive to income dispersion at the lower tail of the distribution show the largest variations in relation to the adoption of different correction methods. Let us compare, for instance, the results for the Atkinson (*e* = 1) index (Part C of Tables 4-5) and Atkinson (*e* = 2) index (Part D of Tables 4-5). The results show that the effects of the correction methods are larger in the case of (*e* = 2). For instance, Table 4 shows that while the range of estimates for the Atkinson (*e* = 1) index reaches 15 percentage points (comparing deletion and two-step regression), it reaches more than 25 percentage points in the case of (*e* = 2). This is a consequence of the coefficients' sensitivity to dispersion at the lower tail, which is differently affected by the distinct methods. Thus, the use of the hot-deck, for instance, increases dispersion at the lower tail and that affects all the coefficients, but it affects relatively more the Atkinson (*e* = 2), which shows the largest variation.

As these simulations show, inequality coefficients can have very different values, depending on the methods used to impute missing incomes. Naturally, the assessment of the true inequality situation (and related concepts, such as poverty or economic welfare) is, under these conditions, a difficult exercise. Inequality evaluations made on the basis of the Gini coefficients, for instance, may differ by up to 7 percentage points if we delete individuals with missing incomes or impute their incomes using a two-step regression model (Table 5, Part A). For the other inequality coefficients, such a discrepancy is even larger. In all cases, the inequality panorama will be highly dependent on the correction method applied, with the additional fact that no method assures the obtaining of accurate inequality coefficients. But even if it is known that a particular method produces accurate estimations of overall inequality extreme cautiousness should be exercised, for instance when analysing inequality more deeply (*i.e.* sub-group inequality analysis). The results in Table 6 (MAPEs for all correction methods) show that even when global inequality measures may be accurately computed, individual incomes are not. For instance, when the two-step regression model produces good inequality estimates (*e.g.* the Gini and the Theil index in Table 4) the imputation errors introduced on average and in absolute values are almost 70% of the actual values. Can a policy-maker worried by the level of income inequality (or poverty) dismiss the fact that some of the lowest incomes are imputed with a substantially higher (or lower) income, artificially improving (or worsening) their condition (for instance, by placing some of them above the poverty line and, for others, increasing the poverty gap)? Even if the overall inequality level is correctly estimated, any social policy aimed at the lower income groups will miss a substantial and relevant portion of (nonrespondent) individuals because of the particular imputation method chosen.

TABLE 6

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE) WHEN DATA
IS MISSING AT THE LOWER TAIL

a) Proportion of income missingness: 7%

Method	1995	1998
Deleting	407.01%	552.50%
Simple regression	103.46%	94.95%
Random regression	132.38%	128.77%
Hotdeck	265.63%	330.52%
Two-step regression	68.29%	68.95%

b) Proportion of income missingness: 12%

Method	1995	1998
Deleting	305.08%	403.27%
Simple regression	86.10%	83.38%
Random regression	115.27%	116.26%
Hotdeck	200.87%	243.71%
Two-step regression	54.12%	56.77%

5.2. Incomes Missing at the Upper Tail of the Distribution

Tables 7 and 8 show that when incomes are missing at the upper tail of the distribution, the correction methods considered here produce underestimated inequality coefficients in all the cases. Two facts can explain this result. The first one is related to the characteristics of income distributions: since income dispersion at the upper tail of the distribution is relatively higher than anywhere else in the income distribution, correction methods that tend to reduce that dispersion, such as standard OLS models, will also reduce inequality coefficients. The second fact is related to the simulation exercise undertaken: to have incomes missing at the upper tail of the distribution implies that imputed values would tend to be lower than the actual ones, as the correction methods applied here (parametric ones, such as the OLS models, or non-parametric ones, such as the hot-deck) use respondent incomes, generally lower than those missing, as a basis to impute incomes. This, in turn, will determine that most inequality coefficients (certainly the ones considered here) will be lower than the actual ones. The particular reaction of each inequality coefficient will, again, depend on the sensitivity of each coefficient to income dispersion at the upper tail of the distribution.

TABLE 7

EFECTS OF NONRESPONSE WHEN INCOME IS MISSING AT THE UPPER TAIL AND THE PROPORTION OF MISSINGNESS IS 7% (*)

A. Gini coefficient			B. Theil coefficient		
Method	1995	1998	Method	1995	1998
Deleting	0.8836	0.8958	Deleting	0.8050	0.8240
Simple regression	0.8820	0.8891	Simple regression	0.7731	0.7887
Random regression	0.9143	0.9253	Random regression	0.8360	0.8647
Hotdeck	0.9277	0.9262	Hotdeck	0.9036	0.8854
Two-step regression	0.9159	0.9076	Two-step regression	0.8228	0.8137
True distribution	0.3881	0.4103	True distribution	0.2931	0.3119

C. Atkinson index (e = 1)			D. Atkinson index (e = 2)		
Method	1995	1998	Method	1995	1998
Deleting	0.8301	0.8497	Deleting	0.8826	0.9031
Simple regression	0.8216	0.8378	Simple regression	0.8843	0.9032
Random regression	0.8681	0.8879	Random regression	0.9100	0.9274
Hotdeck	0.8965	0.8928	Hotdeck	0.9218	0.9274
Two-step regression	0.8689	0.8627	Two-step regression	0.9174	0.9201
True distribution	0.2329	0.2612	True distribution	0.4229	0.4872

Ratios between inequality measure after imputing and the actual one (true distribution). A value higher than one implies an overestimation of the true inequality level, whereas a value lower than one implies an underestimation of this level.
(*) 7% of missingness implies that the missing incomes are located at the highest 14% of the distribution.
Figures in italics are statistically non-significant (95%).

TABLE 8

EFFECTS OF NONRESPONSE WHEN INCOME IS MISSING AT THE UPPER TAIL AND THE PROPORTION OF MISSINGNESS IS 12% (*)

A. Gini coefficient			B. Theil coefficient		
Method	1995	1998	Method	1995	1998
Deleting	0.8975	0.9115	Deleting	0.8558	0.8740
Simple regression	0.8850	0.8923	Simple regression	0.7893	0.8002
Random regression	0.9371	0.9479	Random regression	0.8808	0.9096
Hotdeck	0.9531	0.9521	Hotdeck	0.9763	0.9509
Two-step regression	0.9841	0.9692	Two-step regression	0.9480	0.9184
True distribution	0.3881	0.4103	True distribution	0.2931	0.3119

C. Atkinson index (*e* = 1)			D. Atkinson index (*e* = 2)		
Method	1995	1998	Method	1995	1998
Deleting	0.8549	0.8733	Deleting	0.8901	0.9091
Simple regression	0.8255	0.8396	Simple regression	0.8795	0.8986
Random regression	0.8998	0.9171	Random regression	0.9237	0.9379
Hotdeck	0.9376	0.9299	Hotdeck	0.9388	0.9407
Two-step regression	0.9711	0.9492	Two-step regression	0.9802	0.9706
True distribution	0.2329	0.2612	True distribution	0.4229	0.4872

Ratios between inequality measure after imputing and the actual one (true distribution). A value higher than one implies an overestimation of the true inequality level, whereas a value lower than one implies an underestimation of this level.

(*) 12% of missingness implies that the missing incomes are located at the highest 24% of the distribution.

Figures in italics are statistically non-significant (95%).

Correction methods such as the hot-deck, the two-step regression or the random OLS model perform better as they introduce relatively more variation in imputed incomes than, for instance, the standard OLS model. Table 7 shows that the hot-deck, for instance, produces the most accurate estimations in all cases. However, it still underestimates the Gini coefficient by 7% (see Part A), the Theil coefficient by 11% (Part B), the Atkinson (*e* = 1) by 10% (Part C) and the Atkinson (*e* = 2) by 7% (Part D). The random OLS model and the two-step regression both produce similar estimates, underestimating the Gini by 9%, the Theil by around 17%, the Atkinson (*e* = 1) by 13% and the Atkinson (*e* = 2) by 8%. Finally, the deletion and the standard OLS model produce the least precise estimates. In the case of the Gini and the Atkinson (*e* = 1), the underestimation introduced by these methods is around 12%. In the case of the Theil index, the deletion underestimates the true Theil by around 19%, while the standard OLS does it by around 22%. For the Atkinson (*e* = 1), the underestimation is 16% for the deletion and 17% for the standard OLS model.

Table 8, displaying simulations with a proportion of income missingness of 12%, presents a different panorama. Whereas, for some correction methods, such as deletion or the standard OLS model, the figures obtained are similar to those presented in Table 7 (with 7% of income missingness), for other methods, such as the hot-deck, the random OLS or the two-step regression, the results are more accurate than those presented in Table 7. Thus, the two-step regression, for instance, underestimates the Gini coefficients by around 3% only, the Theil coefficients by 5-8%, the Atkinson (e = 1) by 3-5% and the Atkinson (e = 2) by 2-3%. The reason for this lower underestimation, when the proportion of income missingness increases, is related to the way that simulations are carried out. A low percentage of income missingness means that only very high incomes will be missing (*i.e.* if we simulate 7% of income missingness it means that only the top 14% of the actual incomes have a positive probability of being missing). On the contrary, a higher percentage of income missingness means that a larger area of the income distribution will be affected by the contamination (*i.e.* an income missingness of 12% means that the top 24% of the actual incomes will have a positive probability of being missing). That will imply that, as the proportion of income missingness increases, not only will very high incomes be affected by missingness but also incomes located around the mean of the distribution. Incomes at the middle of the distribution do have different characteristics from incomes located at the top, especially regarding dispersion, as they are relatively more concentrated. Thus, when incomes are missing not only at the top but also in the middle part of the distribution we have the result that certain imputation methods (*e.g.* the hot-deck and the two-step regression) produce two effects: one, by imputing incomes in the top with a lower income, they tend to decrease inequality measures; and two, by introducing income dispersion at the middle of the distribution (an area with relatively less income dispersion than at the top), they tend to increase inequality coefficients. As income missingness increases the second effect dominates the first one, producing an overestimation of inequality coefficients.

From an economic perspective, the results obtained from the simulations may be even more relevant than from a statistical perspective. Incomes located at the top of the actual income distribution are difficult to capture by surveys, as the number of individuals in the population receiving such high incomes is extremely low. It thus becomes crucial not to miss income information from this group when its members are selected in the sample. The main reason is that high incomes have a comparatively larger effect on commonly used inequality measures, such as the Gini coefficient.[6] Missing such incomes would cause an underestimation of inequality, influencing the social debate on it and any corrective policy. This problem (a sampling measurement error) has been documented, at least, in the Latin American case. For instance, Székely and Hilguert (2007) presents maximum incomes registered by many household surveys in Latin America and they compare them with the wage of a typical firm manager. In 10 out of 16 countries analysed, the average of the 10 richest households' incomes was below the wage of a typical middle-sized national firm manager. In Argentina, for instance, the average of the 10 richest households'

[6] This is demonstrated by comparing the effect that the deletion of cases with missing incomes has on the Gini coefficient when considering incomes missing at the bottom of the distribution (Part A of Tables 4 and 5) *vis-à-vis* incomes missing at the top of the distribution (Part A of Tables 7 and 8).

income was 31% *lower* than the earnings of a typical manager. It is likely that errors of this kind are attributable not only to sampling errors, but also to underreporting of incomes at the top level.

TABLE 9

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE) WHEN DATA IS MISSING AT THE UPPER TAIL

a) Proportion of income missingness: 7%

Method	1995	1998
Deleting	59.83%	62.32%
Simple regression	43.75%	43.83%
Random regression	51.27%	50.88%
Hotdeck	60.94%	57.77%
Two-step regression	40.38%	38.83%

b) Proportion of income missingness: 12%

Method	1995	1998
Deleting	49.64%	52.17%
Simple regression	38.55%	38.23%
Random regression	50.32%	50.89%
Hotdeck	58.34%	57.00%
Two-step regression	42.41%	39.52%

VI. CONCLUSIONS

This essay gives an idea of the magnitude and direction of the biases that could be introduced in the measurement of inequality when nonresponse is high. By using several well-known inequality coefficients, each of them sensitive to different dimensions of inequality, it has also been shown how different patterns of missingness can affect such coefficients. The results of the simulations show that all the correction methods considered impute missing incomes with error, which can be large in certain cases and under certain conditions. They also show that in the presence of high nonresponse rates the election of a particular correction method/coefficient could significantly alter the inequality panorama obtained. In the case of nonresponse rates varying in time (or patterns of nonresponse changing in time) it is even possible to obtain inequality estimates that reflect such changes in nonresponse rates (or patterns of nonresponse) rather than in the true inequality situation.

Unfortunately, the simulations show that none of the correction methods considered provide accurate estimates for all the inequality coefficients under all the different patterns of missingness considered. Instead, they show that the use of some methods is not a good idea. For instance, the simple deletion of individuals, the most usual practice in empirical studies, can be a bad strategy unless missingness is randomly allocated across the distribution. If this is not the case, removing individuals from the sample may introduce biases in the measurement of inequality. In such a situation, the use of OLS methods to impute missing incomes (the other preferred method in empirical studies) should also be avoided as this methodology decreases the income variability of imputed incomes and, consequently, may underestimate overall inequality. In this respect, OLS methods should be used with random errors. In the context of the simulations performed in this essay, random OLS proved to be superior to standard OLS in the vast majority of the cases (22 out of 24 cases).

This result suggests that when analysing inequality from data sets containing a significant proportion of missing incomes (such as the Argentinian case) two issues should be considered. The first one is the correct gauging of the pattern of missingness followed by the data. Additional information coming from administrative or tax records should be used whenever possible (as in Atkinson and Micklewright, 1983, for instance), to find out the pattern of missingness, as knowing such a pattern would allow the imputation of missing incomes with a suitable method minimising the risk of obtaining biased inequality coefficients. If there is no secondary source of information from which to infer such a pattern (as in the case of Argentina and most Latin American countries), it should be estimated from the household survey itself. For instance, by estimating a reporting-decision regression, such as the one presented in Section 3, it can be known if there exists a statistical relationship between the probability of income being missing and a set of explanatory variables that are completely recorded in such household surveys. In a case where this reporting-decision regression shows a *non-ignorable* pattern of missingness two-step regression methods or other imputation methods considering this pattern should be used to impute missing incomes.

The second issue that should be considered and that is especially relevant in cases when there is no certainty about the specific pattern of missingness, is the inequality coefficient used to measure inequality. As is clear from the results obtained, the Gini coefficient is relatively less affected by contamination due to nonresponse and its correction. Other coefficients, such as the Theil or the Atkinson indexes, are relatively more vulnerable to contamination, as they can be measured with large errors if the correction methods do not impute incomes accurately in the part of the income distribution where these coefficients are sensitive. Naturally, this imposes a cost on the characterisation that can be made on the distributive situation of a country, as several inequality dimensions cannot be measured as precisely with the Gini coefficient as with other indexes.

Finally, even when certain methods produce accurate inequality estimations, extreme care has to be taken with the inferences made in these situations. Measures of overall inequality may be estimated accurately, but that does not prevent other aspects of inequality, such as subgroups inequality, from being distorted by the use of particular correction methods.

REFERENCES

ALTIMIR, O. and L. BECCARIA (1999). "La distribución del ingreso en la Argentina", *Serie Reformas Económicas* 40, ECLAC, Santiago de Chile.

ATKINSON, A. and J. MICKLEWRIGHT (1983). "On the Reliability of Income Data in the Family Expenditure Survey 1970-1977", *Journal of the Royal Statistical Society*, Series A (General), 146 (1), pp. 33-61.

ATKINSON, A; L. RAINWATER and T. SMEEDING (1995). "Income Distribution in OECD Countries. Evidence from the Luxembourg Income Study", *OECD Social Policy Study* 18.

ATKINSON, A. and F. BOURGUIGNON (2000 a). *Handbook of Income Distribution, Vol. 1*, Elsevier Science.

ATKINSON, A. and F. BOURGUIGNON (2000 b). "Introduction: Income Distribution and Economics" in Atkinson, A. and F. Bourguignon (eds.), *Handbook of Income Distribution*, *Vol. 1*, Elsevier Science.

BANKS, J.; Z. SMITH and M. WAKEFIELD (2002). "The distribution of financial wealth in the UK: Evidence from 2000 BHPS data", The Institute for Fiscal Studies Working Paper WP02/21. London, UK.

BIEWEN, M. (1999). "Item Non-Response and Inequality Measurement: Evidence from the German Earnings Distribution", Discussion Paper 298, Ruprecht-Karls-Universitat, Heidelberg.

BRIGGS, A.; C. TAANE; J. WOLSTENHOLME and P. CLARKE (2003). "Missing… presumed at random: cost-analysis of incomplete data", *Health Economics* 12, pp. 377-392.

CHAMPERNOWNE, D. (1974). "A Comparison of Measures of Inequality of Income Distribution", *The Economic Journal* 84 (336), pp. 787-816.

COWELL, F. (2000). "Measurement of Inequality", in Atkinson, A. and Bourguignon, F. (eds.), *Handbook of Income Distribution* 1, Elsevier Science.

COWELL, F. A.; J. LITCHFIELD and M. MERCADER-PRATS (1999). "Income Inequality Comparisons with Dirty Data: The UK and Spain during the 1980s", *DARP Discussion Paper* 45, LSE-STICERD, London.

COWELL, F. and M. VICTORIA-FESER (1996). "Robustness properties of inequality measures", *Econometrica* 64 (1), pp. 77-101.

COWELL, F. and M. VICTORIA-FESER (2001). "Distributional Dominance with Dirty Data", *DARP Discussion Paper* 51, LSE-STICERD, London.

COWELL, F. and E. FLACHAIRE (2002). "Sensitivity of Inequality Measures to Extreme Values", *DARP Discussion Paper* 60, LSE-STICERD.

DEATON, A. (1997). *The Analysis of Household Surveys*. The World Bank. Washington, DC.

FERES, J. (1998). "Falta de Respuesta a las Preguntas sobre el Ingreso", mimeo. 2do Taller MECOVI. Buenos Aires.

GASPARINI, L. (1999). "Desigualdad en la Distribución del Ingreso y Bienestar. Estimaciones para la Argentina", in *La Distribución del Ingreso en la Argentina*, FIEL. Buenos Aires.

GASPARINI, L. and W. SOSA (1999). "Distribución del Ingreso y Bienestar en la Argentina, 1980-1998", presented at the Annual Meeting of the Asociación Argentina de Economía Política.

GASPARINI, L.; M. MARCHIONNI and W. SOSA (2001). *Distribución del Ingreso en Argentina: perspectivas y efectos sobre el bienestar.* Fundación Arcor - Triunfar S.A. Argentina.

GREENLEES, J.; W. REECE and K. ZIESCHANG (1982). "Imputation of Missing Values When the Probability of Response Depends On the Variable Being Imputed", *Journal of the American Statistical Association* 77 (377), pp. 251-261.

HECKMAN, J. (1979). "Sample Selection Bias as a Specification Error", *Econometrica* 47, pp. 153-161.

KAKWANI, N. and J. SILBER (2008). *Quantitative Approaches to Multidimensional Poverty Measurement.* Palgrave Macmillan.

KING, G.; J. HONAKER; A. JOSEPH and K. SCHEVE (2001). "Analysing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation", *American Political Science Review* 95 (1), pp. 49-69.

LARRAÑAGA, O. (1999). "Distribución de Ingresos y Crecimiento Económico en Chile", *Serie de Reformas Estructurales* 35, ECLAC. Santiago de Chile.

LILLARD, L; J. SMITH and F. WELCH (1986). "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation", *Journal of Political Economy* 94 (3), pp. 489-506.

LITTLE, R. and D. RUBIN (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons.

LITTLE, R. and D. RUBIN (1990). "The Analysis of Social Science Data with Missing Values", in J. Fox and J. Long (Eds.), *Modern Methods of Data Analysis*. Sage Publications.

PAULIN, G. and D. FERRARO (1994). "Imputing income in the Consumer Expenditure Survey", *Monthly Labor Review* 117 (12), pp. 23-31.

RUIZ-TAGLE, J. (1998). "Chile: 40 años de desigualdad de ingresos", mimeo, Departamento de Economía, Universidad de Chile.

SILBER, J. (1999). *Handbook on Income Inequality Measurement.* Kluwer Academic Publishers.

SZEKELY, M. and M. HILGERT (2007). "What's Behind the Inequality We Measure: An Investigation Using Latin American Data", *Oxford Development Studies* 35 (2), pp. 197-217.

VICTORIA-FESER, M. (2000): "Robust Income Distribution Estimation with Missing Data", *Distributional Analysis Research Programme Discussion Paper* 57, London School of Economics, UK.

WEEKS, M. (2001). "Methods of imputation for Missing Data", mimeo, Faculty of Economics and Politics, University of Cambridge.

WOOLDRIDGE, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.