Estimating Inequalities with Missing Incomes.

Paolo Brunori

Pedro Salas-Rojo

Paolo Verme

September, 2022

Estimating Inequality with Missing Incomes

Paolo Brunori[*], Pedro Salas-Rojo[†], and Paolo Verme[‡]

July 26, 2022

Abstract

The measurement of income inequality is affected by missing observations, especially if they are concentrated on the tails of an income distribution. This paper conducts an experiment to test how the different correction methods proposed by the statistical, econometric and machine learning literature address measurement biases of inequality due to item non response. We take a baseline survey and artificially corrupt the data employing several alternative non-linear functions that simulate patterns of income non-response, and show how biased inequality statistics can be when item non-responses are ignored. The comparative assessment of correction methods indicates that most methods are able to partially correct for missing data biases. Sample reweighting based on probabilities on non-response produces inequality estimates quite close to true values in most simulated missing data patterns. Matching and Pareto corrections can also be effective to correct for selected missing data patterns. Other methods, such as Single and Multiple imputations and Machine Learning methods are less effective. A final discussion provides some elements that help explaining these findings.[1]

Keywords: Income Inequality; Item non-response; Income Distributions; Inequality Predictions; Imputations.

JEL Codes: D31; D63; E64; O15.

---

[*]International Inequalities Institute, London School of Economics and University of Bari.

[†]Complutense University of Madrid.

[‡]World Bank.

1 Introduction

Official estimates of inequality worldwide rely on survey data and money-metrics of well-being such as income, consumption, or expenditure. These monetary indicators are known to suffer from unit and item non-response, which are particularly frequent around the tails of income distributions. If left untreated, such phenomenon can potentially bias the measurement of inequality (Rubin, 1983; Solt, 2009; Atkinson et al., 2011; Kennickell, 2017).

A variety of solutions have been proposed to address this problem. The simplest practice consists of ignoring missing observations. Another common approach replaces missing values with means or other moments of the objective distribution. More refined techniques include single and multiple imputations methods (Rubin, 1987); cross-survey imputations (Elbers et al., 2003), replacement of missing observations based on parametric functions (Cowell and Flachaire, 2007; Jenkins, 2017), and sample reweightings based on the probability of non-response (Korinek et al., 2006; Hlasny and Verme, 2018; ?). The increasing adoption of Machine Learning (ML) methods among social scientists (Athey and Imbens, 2019) has recently expanded the portfolio of available options to treat missing data, including regularized regressions and random forest methods (Bertsimas et al., 2018).

As a general norm, the central problem with missing observations consists of knowing whether these observations are missing at random or not (Rubin, 1978). If the missing pattern is completely random, one can safely assume that ignoring these observations does not bias the measurement of inequality. If missing observations are not random, one first needs to understand the underlying missing data pattern and try to correct the data accordingly. This information is typically unavailable to researchers working with surveys, and the methods aimed to address incomplete or corrupted data do not benefit from a counterfactual to test whether they are actually able to correct the missing data problem. It is generally impossible to know whether some methods are actually more or less effective than others in addressing measurement biases due to item non-response because the complete distribution of incomes is unknown.

This paper proposes to address this problem by conducting an experiment focused on income item non-response.[2] We take a real set of income data which we consider clear

---

[2]Note that this paper focuses on incomes and item non-response only. Results cannot be extended to other money-metrics of well-being such as consumption or expenditure, and most correction methods analyzed in this paper cannot be applied to the case of unit non-response.

of missing data and corrupt it mimicking alternative missing data patterns as observed in the literature. We then measure the bias generated by missing observations on the Gini inequality index and test the capacity of correction methods to address this bias. For this purpose, we compare the performance of ten correction methods including deletion of observations, replacement with means, single imputation, multiple imputation, Predictive Means Matching (PMM), replacement with Pareto distribution, reweighting with sample non-responses, and the machine learning methods of LASSO, single tree and random forest.

The paper finds that sample reweighting based on the probability of non-response is the best method to correct income inequality biases due to item non-response. This is true for the most common case encountered with income data, when missing incomes are correlated with income itself and some covariates of income. Even when half of the complete original sample is missing, this method manages to reduce the measurement bias very significantly. It also shows a good performance in most other missing data patterns considered, although its performance declines when missing observations are extremely concentrated at the top. Replacing with means, a common practice among practitioners, is instead the worse performing method in all applications we considered in this paper.

The Pareto-tail adjustment is successful in reducing the inequality bias when missing values are concentrated on the right-tail of the income distribution and, in this respect, is a good complement to the Reweighting method. In our applications, this method seems to overestimate inequality but can potentially be improved with out-of-sample data such as tax data that may help to improve on the parameters of the Pareto function. Predictive Means Matching can also perform relatively well in addressing missing incomes driven by income and other covariates.

Single and multiple imputations reduce the inequality bias when missing values are concentrated in the middle and the left tail of the income distribution, but never as efficiently as sample reweighting, and these missing data patterns are rare cases with income data. Unsurprisingly, we find machine learning methods to be powerful in reducing the root mean squared error between the corrupted and the original data and this can make a real difference when predicting incomes at the individual level. However, under rather general conditions, this comes at the cost of increasing the inequality estimation bias. Based on these findings, we discuss the possible tension between two apparently similar objectives: predicting income at the individual level and predicting inequality for the population at large.

The remainder of the paper is organized as follows. The next section describes the prob-

3

lem of missing incomes with a special focus on different patterns of item non-responses. The third section outlines the most popular methods used in the literature to address the problem of item non-response in surveys. In section four, we propose an empirical application based on a real data set, corrupting the data and applying the methods explained in the previous section. A summary and a discussion of our main results, providing some guidance for practitioners, conclude the paper.

2   The problem of missing incomes

Missing incomes in surveys are mainly caused by item non-responses, unit non-responses and misreportings. Item non-response occurs when individuals participating in surveys do not reply to income questions, while unit non-response refers to individuals who refuse to take part in surveys.[3] It is also possible that individuals taking part in surveys and responding to income questions provide inaccurate information on incomes. When this is likely, as in the case of negative or tiny wages, statistical agencies or researchers may prefer to omit this information from the original sample, hence enlarging the missing data problem.

In this paper we focus on income item non-response. This problem is less restrictive than unit non-response, because the remaining information available in the survey for individuals not responding to income questions can be used to estimate income with predictive methods such as single or multiple imputation models (see section 3). Focusing on item non-response has two advantages: it enables the comparison of a larger set of correction methods as compared to methods designed for unit non-response while conclusions derived from item non-response corrections would also apply to correction methods designed to address unit non-response issues.

Statisticians refer to three types of missing data patterns that, in the case of income, can be described as follows (Rubin, 1978):

- Missing Completely at Random (MCAR) - when the probability of missing incomes is the same across the full distribution of incomes;

- Missing at Random (MAR) - when the probability of missing incomes depends on non-income observed variables;

---

[3]Note that we always refer to individuals' non responses, although all ideas expressed in this paper would also apply to households' non responses.

- Missing Not at Random (MNAR) - when the probability of missing incomes depends on income itself.

This classification is relevant because it ultimately determines whether missing data are problematic or not, and the type of statistical problem to address. MCAR data is evidently the least damaging. If one can argue that missing observations are MCAR for all income and non-income variables, estimating statistics from the remaining observed sample is a viable option with the only caveat that weights should be adjusted to provide accurate population figures. This missing pattern is considered to be ignorable (Rubin, 1976).

If missing data are MAR, providing statistics by non-income variables may be problematic. For example, suppose that missing income observations are random across the income variable but non-random across gender, with all missing income observations being females. In this case, using gender in an income regression or providing income inequality statistics by gender groups could deliver biased results (unless missing incomes are distributed randomly within each group).

Finally, if missing income values are MNAR, the missing pattern is defined as a function of incomes. This situation is a well-known attribute of incomes reported in surveys worldwide: individuals with higher incomes are less likely to respond income questions and more likely to under report incomes (Jenkins, 2017), and there is also recent evidence suggesting that lower income individuals are less likely to respond to income questions (Hlasny et al., 2021).

Clearly, MNAR is the most problematic of the three cases considered, because one first needs to understand the function that defines missing data before applying any correction method. Although the general assumption with income is that higher income individuals have a lower probability of response, the precise probabilities are usually unknown to practitioners. MAR data are also very likely to be MNAR. Indeed, the case of missing incomes correlated with income and some covariates is the most common case encountered by practitioners working with incomes. Note that this is what we will label "MAR" in the remaining of the paper.

In practice, it is not easy to determine the missing data pattern. The MCAR pattern can be tested with the Little's test (Little, 1988), which is used to check whether the null hypothesis of missing observations being completely at random is rejected or not. The MAR pattern can be assessed by testing the correlation of observed incomes with non-income variables, which provides indirect information on the correlation between non-observed

incomes and non-income variables. But the MNAR pattern is unknown unless one has access to reliable complementary out-of-survey data such as tax or administrative data[4], or other in-survey data that can be used to estimate the probability of income non-response such as non-response rates by administrative area.

The case of income inequality measurement is also particularly problematic if data are MNAR. Several authors who worked on the US Population Census noted that income non-responses are U-shaped, with lower and higher income individuals being less likely to respond to income questions (Rubin, 1983; Greenlees et al., 1982). Scholars who work on top incomes routinely find that missing incomes increase with income, which means that missing observations grow in number and mass as we approach the upper tail of the income distribution (Atkinson et al., 2011, Jenkins, 2017). We also know that higher incomes weigh more on inequality measures than lower incomes (Ceriani and Verme, 2021). These stylized facts vary depending on the inequality measure considered but they generally apply to most inequality measures (Cowell and Flachaire, 2007). Moreover, missing incomes on the lower tail can also be frequent and problematic for the measurement of inequality, although their number and mass make them less problematic than missing high incomes (see Ceriani and Verme, 2021 and Hlasny et al., 2021).

In the remaining of the paper, we will focus on the Gini index of inequality. This choice is driven by its popularity among income inequality specialists, and by the fact that this index attributes more weight to observations located in the middle of the income distribution as opposed to other measures of inequality, such as the Theil index or the top-bottom income share (Cowell and Flachaire, 2007). The Gini index is, therefore, less sensitive to missing observations in the tails of the income distribution than other indexes of inequality, and we would expect the bias generated by missing incomes on the tails to be on the lower side as compared to other measures. We will report basic statistics on the Mean Logarithmic Deviation (MLD), the Atkinson Index (ϵ=2), 20% share, 80% share and the 20%/80% ratio whereas full results on these indexes are available on request.

---

[4]One should always be cautious when using tax or administrative data, because tax exemptions, non-filers or tax avoidance could also affect the variable under scrutiny, hence biasing the estimates aimed to cover the complete population.

3 Popular solutions to address missing data issues

In this section we explain the most popular methods employed in statistics and econometrics to address issues related to item non-response. By order of exposition, we explain deletion, replacing with means, replacing with parametric imputations -including single and multiple imputations-, replacing with matching, replacing with parametric distributions, reweighting techniques and machine learning models including LASSO, conditional inference trees and random forests.[5]

As a baseline set up, consider individual j in a population of n individuals characterized by a vector of characteristics $x_j^1, ..., x_j^C$ with income y_j. The objective is to estimate the Gini inequality index on the full distribution of incomes from a survey that is assumed to be representative of the population of interest. Data on non-income variables are observed in full while a share of income observations M is missing. For simplicity, we assume that all observations have the same sample weight. We also assume that the survey design accurately reflects the objective population, such that the complete data delivers non-biased estimates of income inequality.

3.1 Deletion

Deletion consists in ignoring missing incomes or erasing incomes with suspected misreported information. Inequality is estimated only on observed values, without applying any other correction. The main advantage of this approach is its simplicity, making it a popular shortcut for practitioners. In principle, in the context of income inequality, if missing data are MCAR, deletion should deliver inequality estimates in expectation identical to those expected from the full distribution, although the standard errors may be large. If missing data are MAR, estimates of overall inequality are expected to be similar to those from the complete distribution. If missing data are MNAR, deletion results invariably in biased inequality measures.

3.2 Replacing with means

This method consists of replacing missing observations with the average of the remaining observed values.

---

[5]Table A2 in the Appendix contains the software packages used to implement the different correction methods.

$$\forall i \in M : y_i = \sum_{j \notin M} w_j y_j \tag{1}$$

where w_j is estimated over the number of non-missing incomes.

This replacement can result from averaging values based on geographical sub-sub-groups, such as primary sample units, administrative areas or the full population average, or sometimes averaging values by covariates of income. The main advantage of this method is, again, its simplicity. However, replacing missing incomes with the mean reduces income variability. As compared to the deletion approach, replacing missing observations might improve estimations of the standard error for MCAR and MAR data, but the potential inequality biases generated by MAR and MNAR data will persist and would increase. Moreover, in right-skewed distributions such as that of income, the mean is by definition higher than the median, so this imputation method tends to overestimate imputed incomes in the left tail.

3.3 Replacing with parametric imputations

Replacing with predicted values or imputations consists in substituting missing income observations with values generated with a regression model. The two most common tools are based on single and multiple parametric imputation methods. The simplest case is **single imputation** described as follows:

$$\forall j \notin M : log(y_j) = \boldsymbol{\beta} \mathbf{x}_j + \epsilon_j \tag{2}$$

$$\epsilon_j \sim N(0, \sigma^2) \tag{3}$$

where $\mathbf{x}_j$ is the vector of observed covariates, $\boldsymbol{\beta}$ is the vector of unknown regression coefficients to be estimated and σ^2 represents the variance of the error term ϵ_j. The missing incomes are then imputed by fitting the parameters as follows:[6]

---

[6]To obtain non-logarithmic incomes, it is not enough to apply the transformation $exp(\hat{\boldsymbol{\beta}} \mathbf{x}_i)$. Log-Linear regression models deliver consistent estimates only under the -potentially strong- assumption that the distribution of the error term is independent of the regressors. As shown in Blackburn, 2007, this assumption is violated by the presence of heteroskedasticity, which is partially but not completely corrected by robust standard errors. Hence, to obtain more reliable predicted values, especially on the right tail, it is necessary to apply an extra transformation: $\hat{y}_i = exp(\hat{\boldsymbol{\beta}} \mathbf{x}_i + \sigma^2/2)$.

$$\forall i \in M : \hat{y}_i = exp(\hat{\boldsymbol{\beta}}\mathbf{x}_i + \sigma^2/2) \tag{4}$$

With sample surveys, single imputation may deliver accurate means but the predicted variance is expected to be inaccurate as the model relies on one draw of data. In this case, it is usually recommended to draw several bootstrapped samples that somehow reflect the potentially different realizations of the missing pattern. This is what is referred to as **multiple imputation** (Rubin, 1987, Rubin, 1996 or Raghunathan et al., 2001).

Single or multiple parametric imputation methods can be applied to MCAR and MAR missing income data, but they are not able to address issues generated by MNAR data. To do so, one should know the function regulating non-response probabilities. Rubin, 1983 discusses data on wages from the US Current Population Survey and argues that multiple imputation cannot address biases generated by MNAR data because the function regulating the missing data pattern is unknown and the prediction of missing data is unlikely to be accurate. This method is expected to improve estimations of the standard errors but cannot correct for bias inequality estimates.

3.4 Replacing with matching

Another common approach to imputing missing observations is based on matching methods. The general idea consists of finding some observations within the observed sample that potentially match as closely as possible observations with missing incomes. The matching process is based on the remaining observed covariates, which can be regarded as predictors of income. The matching literature is rich and multiple matching methods have been proposed to address missing data issues (Stuart, 2010). One drawback of these methods is that they generally assume that missing observations are unrelated to income (Lillard et al., 1986). A priori, they should function well for MCAR and MAR, but not for MNAR data.

In this paper, we use the Predictive Means Matching (PMM) method proposed by Rubin, 1986 and Schenker and Taylor, 1996. First, we define the predictive mean of a missing observation as in Equation 4. Then, instead of assigning the predicted income to each incomplete income case, we draw a random observation from the set of complete cases - which are considered as donors- having the predictive means coinciding with the incomplete case. The correspondent value of y_j of the selected observation is then imputed in the incomplete case.

3.5 Replacing with parametric distributions

This method consists of replacing observed values in selected parts of an income distribution with other values extracted from a theoretical parametric distribution. In the past decades, the inequality literature has proposed several functional forms to replace observations on both tails of an income distribution (see Cowell and Flachaire, 2007 and Jenkins, 2017). Replacing missing incomes with values extracted from a parametric distribution is suitable when missing incomes are due to either item or unit non-response, can potentially address all types of missing data patterns, and are particulalry effective when missing observations are concentrated on the tails. Their limitation reside in the fact that they require a cutpoint below or above which observations are not replaced. If missing incomes are present all along the distribution, these methods are not very effective in correcting biases generated by missing observations.

In this paper we lean towards the most common approach: the Pareto right-tail distribution adjustment. This adjustment consists of replacing values above a certain threshold to make it fit the shape of a Pareto-tail distribution. This distribution is usually defined in terms of its own cumulative distribution function, such that:

$$F_\theta(y) = 1 - \left(\frac{y_j}{y_0} \right)^\theta, y_j \geq y_0 \tag{5}$$

where $y_0 > 0$ is a scale parameter defining the threshold over which the Pareto tail is adjusted, and $\theta > 0$ is a data-specific shape parameter.[7]

3.6 Reweighting

Reweighting methods adjust the weights of non-missing observations in order to correct the statistics of interest. In other words, rather than replacing missing observations and expanding the sample, these methods simply adjust the weights of observed incomes to account for missing observations.

The reweighting method used in this paper is the one proposed by Korinek et al. (2006)

---

[7]We estimate the shape parameter using the integrated squared error (ISE) estimator proposed by (Vandewalle et al., 2007), that can be written as:

$$\hat{\theta} = argmin_\theta \left[\int f_\theta^2(y) dy - 2\mathbb{E}(f_\theta(y)) \right] \tag{6}$$

where $f_\theta(y)$ is the density function of the Pareto distribution for the relative excesses.

and used in various applications by Hlasny and Verme (2018), **?**, and Munoz and Morelli (2021a). This is a probabilistic model that uses non-response rates across geographical units to estimate the probability of non-response. Following **?**, the probability of non-response is defined as

$$P_i(x_i, \theta) = \frac{e^{g(x_i, \theta)}}{1 + e^{g(x_i, \theta)}} \tag{7}$$

where $g(x_i, \theta)$ is a stable function of observable characteristic x and θ is a vector of parameters estimated with a Generalized Method of Moments (GMM) model as:

$$\hat{\theta} = argmin \sum_{j=1,\dots,J} \left[(\hat{m}_j - m_j) w_j^{-1} (\hat{m}_j - m_j) \right] \tag{8}$$

where m_j is the number of households in region j according to sample design, $\hat{m}_j$ is the estimated number of households in the region, and w_j is a region-specific analytical weight proportional to m_j. The estimated number of households $\hat{m}_j$ can be imputed as the sum of inverted estimated response probabilities of responding households in the region $\hat{P}_{ij}$ where the summation is over all n_j (population for area j) responding households.

To the best of our knowledge, this is the only method that explicitly learns from the missing data pattern, determines the function that explains the probability of non-response, and uses this function to address issues related to MAR and MNAR data. Its drawback is that one requires information on non-response rates by geographical areas, which is not always available to researchers.

3.7 Machine Learning Models

Supervised machine learning can also be used to handle item non response. Similarly to standard multiple imputation methods, these algorithms are trained on a joint distribution of an output variable (income) and several input variables (covariates), to obtain the best possible out-of-sample prediction of an output value. Machine learning algorithms are particularly suitable to detect non-linearities in the data generating process, as well as to perform variable selection to avoid model overfitting when the number of covariates that can potentially be used to impute missing observations is large (Varian, 2014; Athey and Imbens, 2019). In this paper, we limit the use of ML techniques to three of the most popular methods used in economics: regression with regularization, trees and random forests.

Regressions with regularization. Regularization methods are used to reduce the risk of overfitting in linear regressions, hence improving their prediction accuracy. Standard regression parameters are shrunk by applying a penalization term that reduces their absolute value. The optimal shrinkage is determined by assessing the out-of-sample mean squared error of alternative values of the penalization term.

The most popular regularization algorihtm is the LASSO (Least Absolute Selection and Shrinkage Operator, introduced by (Tibshirani, 1996). LASSO can be formally defined as a regression where the coefficients of Equation 2 are estimated by minimizing:

$$\forall j \notin M : \sum_{j=1}^{n} \left(log(y_j) - \boldsymbol{\beta}\mathbf{x}_j\right)^2 - \lambda \sum_{c=1}^{C} |\beta_c| \tag{9}$$

where the first term is the typical sum of squared residuals and λ is the penalization term. When $\lambda = 0$, LASSO is equivalent to an OLS, while larger λs result in more shrunk coefficients. Eventually, for a sufficiently large λ, one or more coefficients will be set to zero, so the algorithm simultaneously regularizes and performs variable selection. Among alternative penalization terms, the choice is guided by an estimate of the out-of-sample predictive accuracy of the resulting model (Tibshirani, 1996; Hastie et al., 2009). Once the LASSO parameters are fitted, missing incomes are predicted as in Equation 4.

Decision trees and random forests. A regression tree is an algorithm trained to predict an output variable out-of-sample by splitting the observed sample into non-overlapping subgroups, based on a partition of the regressors' space. Then, the prediction is typically the average realization of the output variable by subgroups. The binary splitting used to grow trees makes them particularly suitable to fit highly non-linear data generating processes (Hastie et al., 2009; Hothorn et al., 2006). Similar to other prediction methods, we first build the tree estructure by training the subsample of complete observations, and then we predict incomes over missing observations based on their own covariates.

The low bias of trees comes at the cost of a substantial variance that can be reduced by bootstrap aggregation or random forests (Ho, 1995; Breiman, Breiman). Random forest models derive their predictions from analyzing a multitude of decision trees. Each tree is built from a random sample extracted from the original population, and the combination of covariates used to determine each split is also a random set of variables extracted from the complete set available. The idea is to pool multiple decision tree-votes, each one endowed with a different set of information extracted at random from the original data. The final

prediction is produced by averaging all predictions produced by each tree.

4   Empirical analysis

4.1   Data

Simulations are based on the 5th wave of the National Income Dynamics Study (NIDS), carried out in 2017 by the Southern Africa Labour and Development Research Unit (SAL-DRU), at the University of Cape Town. NIDS is one of the most accurate survey on income among low and middle-income countries, has a very low non-response rate relative to other countries, and covers a country characterized by a very high level of income inequality (Schotte et al., 2022), three features particularity useful for our analysis.

The income variable we use is monthly take-home wages from primary and secondary employment (variable $w5\_fwag$). This choice allows us to limit issues of item non-response and misreporting to a minimum.[8] In the original data, there are 7,591 observations with labor incomes and a complete set of covariates, with 7,199 originally reported by the respondents and 377 imputed by SALDRU. This leads to an item non-response rate of about 5%, a very low rate by global standards (Riphahn and Serfling (2005)). The Gini estimated on all observations (with imputed incomes) is estimated at 53.1% as compared to a Gini of 53.4% estimated on observed income with overlapping confidence intervals (100 bootstrap repetitions). The Gini estimated on the raw sample (no bootstrap) is 53.2% with imputations and 53.5% without. All estimations are conducted with sample weights.

4.2   Data corruption

For the data corruption exercise and the consequent analysis we assume the complete observed data (7,199 observations) to be complete and untainted by missing data. As a consequence, our analysis should be understood as a laboratory experiment to test the validity of different correction methods and none of the inequality statistics presented in this paper should be considered as accurate representation of inequality statistics in South-Africa.

The essence of our missing data pattern simulation is the attribution of a probability of non-response to each observation. The first determinant of such probability is the share

---

[8] All labor incomes in NIDS are positive and we do not have issues related to imputation of negative incomes.

of missing observations in the complete population. Although we tested many different shares of missing, our final results are illustrated with shares of 3%, 5%, 7%, 10%, 15%, 20%, 30%, 40% and 50%. Showing other shares does not meaningfully alter any of our conclusions. When simulating the MCAR pattern explained in Section 2, the proportion of missing data is the only variable determining the probability of being missing, which is orthogonal to income and any other covariate. This probability is homogeneous to all observations and directly corresponds to the desired share of missing data.

In the case of MAR simulations, the probability of non-response depends on the covariates used as predictors of income. In order to obtain a MAR pattern consistent with South African missing data, we estimate a score vector based on the original data. We go back to the original NIDS sample, including the complete and incomplete labor incomes, and observe the covariates associated to individuals with missing incomes. In particular, we consider all covariates used by SALDRU to impute employed labor income including sex, race, age, age squared, union membership, years of schooling, years of schooling squared, several tertiary education degrees, marital status, number of rooms, number of rooms squared, geotype of the dwelling, month of interview and province of residence (see Argent, 2009 for more information regarding these variables). We then run a logistic regression with the dependent variable being 1 if incomes are missing and 0 if incomes are non-missing, hence exploring the relation between the probability of non-response and the selected covariates. To avoid overfitting, we regularize the logistic regression estimating a LASSO, as explained in Section 3.[9]

Finally, for MNAR simulations, probabilities of non-response are designed to be correlated solely with income. MNAR patterns are simulated following the non-linear functions proposed in Schouten et al. (2018). We consider four possible MNAR patterns: i) the probability of non-response is higher in the middle of the distribution (MID); ii) the probability of non-response is monotonically decreasing with income (LEFT); iii) The probability of non-response is monotonically increasing with income (RIGHT); and iv) the probability of non-response is higher in both tails of the distribution (TAIL). As explained, cases iii) and iv) are the most common cases with income data. However, it is also useful to illustrate results for cases i) and ii) as benchmarks and because these patterns could apply to certain income definitions. Figure 1 shows how the four non-linear probability functions of non-response vary along the distribution of income.

---

[9]The results of this LASSO regression are shown in Table A1, while Figure A1 in the Appendix reports the LASSO tuning output.

We consider the MAR simulation to be the more realistic exercise because missing incomes usually do not depend neither exclusively on a random parameter (MCAR) nor on the sole income distribution (MNAR). In the MAR pattern, factors such as gender, education, the province of residence, or the type of dwelling drive the missingness probability. In our case, we find this association leading to a pattern similar to MNAR/RIGHT, although the MAR covariate-driven approach delivers a smoother missing distribution than that proposed in Figure 1. In other words and in reality, a pure MAR pattern where missing observations are related to a covariate but not to income is extremely rare. With this in mind, our main results and discussions are based on MAR corrections where missing observations are correlated to both income and covariates of income. We will also comment on other data corruption patterns providing results in Appendix.[10]

[FIGURE 1]

4.3   Inequality measurement with data corrections

We now turn to compare the capacity of missing data correction methods to improve on inequality estimates. As a general criterion for the evaluation, we use the arithmetic difference between inequality measured on the corrupted incomes and inequality measured on the original complete incomes and apply a difference-in-means t-test with 100 bootstraped repetitions. In all Tables below, a negative sign is interpreted as an underestimation of inequality estimated with corrupted data, and a positive sign as an overestimation. For all methods based on an explicit imputation of missing values (all but deletion and reweighting) we also provide the Root Mean Squared Error (RMSE) as an estimate of the precision of the model.

One may observe trade-offs between minimizing inequality prediction errors and the RMSE. Suppose that ignoring missing incomes reduces inequality and that this problem is addressed with an OLS imputation model that is able to reduce the downward bias. Now, assume that suspecting the OLS model to overfit the data, it is regularized with a LASSO model, as in Equation 9. The optimal shrinkage is determined by choosing the value of the penalization term producing the smallest out-of sample RMSE. Different values for the penalization term are tested and a lower out-of-sample RSME is obtained. In the general case in which regularization will shrink towards zero the values of the OLS parameters, it will also shrink the variability of predicted incomes, producing a more severely downward

---

[10]More graphical information regarding the generated MAR pattern are available upon request.

biased inequality estimate than what obtained using an OLS imputation. That is to say, if we are downward biased in estimating inequality, worsening individual incomes' prediction including some noise due to overfitting would improve our inequality estimation at the cost of a higher prediction error at the individual level, an issue that emerges for several data correction patterns.

The benchmark statistics obtained with the complete sample are shown in Table 1. Mean monthly labor income amounts to 8,057 rands (2017), with a Gini reaching 53.5%, a MLD index of 0.521 and an Atkinson of 0.689. Around 3.5% of labor income belongs to the bottom quintile, while almost 60% is owned by the top 20%. The p80p20 ratio is 16.87. As expected, these baseline inequality measures indicate a rather extreme level of inequality across labor earnings.

[TABLE 1]

In the reminder of this section, we compare the performance of the correction methods described in section 3 applied over the different missing data patterns described in section 4.2. The main results are reported in Tables ?? to ??. Each table refers to a different missing data pattern and each column in tables refer to a different correction method. Top panels report the statistical difference between the Gini estimated on the corrupted data and the Gini estimated on the original complete distribution of incomes (Gini diff.). The bottom panels report RMSE values for the correction methods where RMSEs can be estimated.

4.3.1 MCAR

As explained, MCAR is the least problematic case but it is important to understand that the Gini bias generated by MCAR may not be zero, and that practitioners are usually blind in regards to the missing data pattern. They may be applying corrections when, in fact, missing data are randomly distributed and not problematic.

Table ?? reports results for MCAR data. As expected, ignoring missing incomes does not lead to a significant Gini bias (all the Gini difference values in the Deletion column are close to zero and non-significant). However, it should be remarked that this only occurs after several repetitions of the bootstrap exercise, more than 40. For a small number of repetitions, MCAR inequality estimates are significantly (and randomly) different from those used as benchmark in Table 1. This is an important finding, because even missing

items thought to be MCAR may result in inequality estimates that do not match true inequality if inequality is simply calculated on the raw sample.

Moreover, if one is blind with respect to the missing data pattern and applies corrections, results show that the Gini becomes biased. That is the case for all correction methods with the exception of reweighting, which is explained by the fact that reweighting is the only method that includes the estimation of the missing data pattern (the probability of non-response), which is then used to correct incomes. In other words, this method first learns from the data the specific missing data pattern, and then applies corrections using this information. Indeed, the reweight column in Table **??** shows that applying this method results in Ginis that are not significantly different from the Gini estimated on complete data, and this is true for any share of missing incomes.

[TABLE 2]

4.3.2 MAR

As explained, the MAR pattern we simulate results in the majority of incomes missing from the middle and top of the income distribution implying a correlation between missing and observed incomes (and between regressors of income and missing incomes), as in the MNAR case. This is the most common and problematic case for practitioners working on the estimation of inequality with incomes.

Table **??** shows results for the MAR missing data pattern. As expected, with MAR data, when missing incomes are ignored the Gini is underestimated. Indeed, the Deletion column (top panel) shows that the Gini is significantly underestimated and the size of the underestimation grows with the share of missing incomes.

Yet, attempting to correct for this bias may not result in better Gini estimates than simply ignoring missing incomes. This is evident looking at the Gini difference results across correction methods. None of the methods is able to eliminate the bias completely. Some methods such as PMM, Pareto and Reweighting do better than Deletion, but other methods such as replacement with Means, Single and Multiple Imputation, and ML methods (LASSO, Tree and Random Forest) do worse. With the exception of Pareto, the poor performance of these methods also becomes worse as the share of missing incomes increases.

More in details, inequality measures obtained after replacing with a *Single Imputation* are underestimated. The bias is lower than for Deletion with lower shares of missing observations but higher for higher shares. However, the correction bias is not as large

17

as that found for the imputation of the mean. For *Multiple Imputation*, we find worse results in that the bias is larger and is always worse than Deletion. This is not surprising, because re-sampling with multiple imputations improves the estimation of the standard error, but tends to worsen the precision of the inequality estimates. The intuition is that the distribution of the average of five predictions has a lower variance than the distribution of the single predictions. If using single imputation induces a downward bias in the prediction, averaging across imputations may worsen the problem.

Predictive Means Matching (PMM) substitutes missing incomes with a random value drawn from the complete set of observations with actual incomes close to fitted incomes in the missing observations. PMM reflects the uncertainty of imputations and does not necessarily impute the same value to missing observations that share covariates. Consequently, correcting with PMM never leads to smaller estimates of inequality than single or multiple parametric imputations. Results in Table **??** confirm that the PMM method reduces more the inequality bias than other parametric imputations. However, the RMSE estimation is larger than that obtained with single and multiple parametric imputations. This result illustrates the trade-off between imputation accuracy and precision in the inequality measurement we have previously explained. Overfitting imputed values increases inequality estimates, because it induces more heterogeneity in the income vector (randomly matching real values conditioned on the fitted means) at the cost of increasing RMSEs. When the objective is simply measuring inequality, and it is irrelevant whether each income is estimated correctly, PMM appears a better approach than simple or multiple parametric imputations, but if the aim is to obtain reliable predictions for missing data, this method is among the worst performing of all.

With the *Pareto* correction, we modify the right tail of the corrupted dataset and make it fit a Pareto distribution. Given the Pareto distribution we have specified,[11] we obtain upward biased estimates of inequality with no clear pattern as the share of missing observations increases. Note that, with this method, we impose a specific parametric shape

---

[11]As explained in Section 3, the Pareto-tail correction is characterized by two parameters: the percentile threshold above which the right tail is fitted and the shape parameter. Regarding the former, we have checked our results with many different thresholds, using the top 10th, 7th, 5th, 2nd and 1st percentiles. Here we lean towards using the top 10 percentile because it is the threshold usually delivering least biased estimates when missing shares are above 15%. Hence, our results can be considered as lower-bound estimates of the inequality bias, although the precision of other thresholds varies depending on the missing pattern and the share of missing under scrutiny. Overall, as explained by (Jenkins, 2017), administrative data is required to obtain data-driven threshold parameters. We have also checked the Hill Estimator and the Partial Density Component Estimator, finding no meaningful differences across their results.

on a certain part of the distribution. For different missing shares, the corrupted distribution will be different and the adjustment of the Pareto correction can be "better" or "worse".

Results for sample *Reweighting* show a rather limited downward bias which grows with the share of missing incomes. This method is the most successful among those considered in correcting for MAR data. In the worst-case scenario, after turning 50% of the sample to missing, the Gini index is underestimated by only 1.6 points, the same bias obtained using multiple imputation with a missing share of 7%. This performance is explained by two factors. One is that this is the only method that estimates the probability of non-response and uses this probability to correct the data. Recall that our MAR data pattern includes the MNAR feature in that missing incomes are correlated also with incomes. The lack of knowledge on the function that relates missing incomes to observed incomes is the reason why multiple imputation cannot address biases generated by MNAR data (Rubin, 1983). This method addresses this problem. The second factor is that this method uses the variance of incomes within groups and between groups to estimate the variance of the full distribution of incomes (observed and unobserved values) unlike other methods such as OLS and machine learning methods that predict incomes using only the variance of observed values.

The idea of the *LASSO* imputation is similar to the single imputation parametric method, but now including the extra regularization component (see Equation 9). Taking advantage of this regularization, we could include many more regressors and interactions, hoping to obtain results closer to the original income distribution. Besides the covariates employed in the other imputation methods, we also include all pairwise interactions between gender (binary), race (white, coloured, asian/indian, african) and the years of schooling. Results in Table **??** show that, with LASSO, the inequality bias is similar to other parametric imputation methods. Including the regularization term and the interactions does not seem to provide a clear improvement in this case. However, in Table **??**, we find the RMSE in LASSO to be slightly smaller than in the other parametric imputations. We find once more the tension between the closeness of the imputation adjustment and the inequality bias.

The same idea applies to *Tree* and *Random Forests (RF)*. In both cases, the accuracy of the imputation is much higher than for other methods, especially for random forests, which in Table **??** shows the smallest RMSEs. However and contrary to the PMM results, this accuracy in predicting incomes comes at the cost of increasing the inequality bias. As the share of missing observations increases, these methods provide an increasingly larger

19

bias becoming the worse performing methods together with Mean replacement as the share of missing observations nears 50%.

[TABLE 3]

4.3.3 MNAR

Although MAR may be the most commonly encountered pattern of missing incomes in real data, there might be some cases in which missing data are MNAR only (missing incomes are only correlated with incomes and not with other covariates). Tables **??** to **??** show MNAR results with different flavors (missing data concentrated in the middle, left, right and both tails of the income distribution). To keep the exposition simple, we focus the discussion on the inequality bias results. RMSEs results are reported in the bottom panels of all tables.

Regarding the MNAR/MID pattern (Table **??**), missing items are more concentrated around the median, so the income distribution becomes more polarized with the increasing shares of missing observations. This provokes a positive sign in the *Deletion* column, because inequality is overestimated in the corrupted data. Similarly, *Pareto* and *Reweighting* produce upward biased estimates. Methods that use covariates to correct for missing data perform better in this case, given the substantial degree of correlation between income and its covariates. The imputation using the *Mean* remains the worst performing method, even if missing items are concentrated around the mean. Single, multiple imputations and the three machine learning methods reduce significantly the inequality bias, with LASSO being especially successful for this task. MNAR-MID is the only missing data pattern where reweighting is outperformed by the majority of the other correction methods. However, while possible, it is unlikely with survey data to have the majority of missing data concentrated in the middle of the income distribution.

Table **??** shows results for the MNAR/LEFT pattern. It is not obvious which effect should be expected on inequality after corrupting the bottom of an income distribution. On the one hand, removing small incomes would make lower quantiles relatively richer (decreasing inequality). On the other hand, top quantiles will lose relatively less affluent individuals, making quantiles at the top even richer (increasing inequality). The net effect depends on the specific shape of the income distribution. In the case of South Africa, the simulated missing pattern mildly increases inequality (see the *Deletion* column). Among correction methods, *PMM* and *Reweighting* are the best performing methods. The for-

mer producing slightly downward biased estimates and the latter overestimated levels of inequality.

Table **??** and **??** show the MNAR/RIGHT and MNAR/TAILS results. As we should expect given our MAR pattern, these results are similar to those obtained with MAR, but more pronounced. In the extreme case, with 50% of the sample missing, both MNAR/RIGHT and MNAR/TAIL provoke an underestimation of almost 11 Gini points (see the *Deletion* column). Interestingly, being both patterns different in the sense that one focuses on top incomes and the other on both tails, the effect on inequality is almost the same, highlighting the importance of the right tail in the inequality bias. In both cases, no model is successful in correcting inequality estimates, although the Pareto adjustment seems to deliver relatively more precise results, which is expected given its focus on top incomes. Also, *Deletion* performs better than the majority of the other methods with the exception of *Pareto* and *Reweighting*.

It is noticeable that *Reweighting* performs well relatively to *Pareto* for low shares of missing incomes but not for high shares. The performance of this method is also much worse than its performance on MAR data. This may be due to two factors. The MNAR/RIGHT and MNAR/TAILS data are much more skewed in income than the MAR data by design. This means that, when Reweighting estimates the probability of non-response, it has little information to use at the very top of the distribution and is likely to underestimate this probability for top incomes. This would be the case, for example, if all households in Primary Sample Units (PSU) with very high incomes do not respond to the income question. It may also be the case that *Reweighting* is able to address correlations of missing incomes with both income and other covariates, whereas *Pareto* is effective in addressing correlation of missing incomes with income but not with covariates of income.

[TABLES 4-7]

4.4 Summary, conclusions, and recommendations

The paper provided a laboratory experiment to test which of the most popular correction methods for income item non-response in surveys provides the best outcomes in the context of income inequality measurement. We used a real data set (NIDS, from South Africa 2017) complete of incomes and relevant covariates and corrupted it mimicking known missing data patterns including MCAR, MAR, and MNAR with various flavors and shares of missing incomes. We then proposed to compare the performance of ten different methods used by

21

statistical agencies, statisticians and economists to correct measurement biases due to item non-response using the Gini index as statistics of interest.

Table 8 summarizes results. We report the arithmetic average of the Gini bias (the mean difference between the Gini measured on corrupted data and the Gini measured on complete data) across shares of missing incomes and for all missing data patterns, and the average RMSE. Note that the RMSE cannot be calculated for the Deletion and Reweighting methods. In terms of the Gini difference, the Reweighting method is the one that performs relative better overall and is the best performing method for the most common MAR case. Replacement with Mean also stands out as the worst performing method. PMM also performs well overall whereas Pareto performs better than other methods for MNAR/RIGH and MNAR-TAILS data. The other methods including Single and Multiple Imputation, LASSO, Tree and Random Forest perform well only with MNAR/LEFT or MNAR-MID data, which are very rare cases with income data.

[TABLE 8]

Figures 2 and 3 summarize main results for the MAR correction. Figure 2 shows the different correction methods in their ability to estimate the Gini correctly as the share of missing observations increases. It can be seen that the decay in prediction accuracy varies across methods. Reweighting is the method that is closer to the true Gini and remains so as the share of missing incomes increases, whereas mean imputation is the method that decays faster than all others. Other relatively good performing methods in this respect are Deletion and PMM. Also noticeable is the fact that Pareto overestimates rather than underestimate inequality and that this method does not show any pattern as the share of missing incomes increases. Figure 3 shows that RMSE increases as the share of missing incomes increases as we should expect but, again, the size and gradient of this increase is different across methods. RMSE growth is the steepest for PMM and the least steep for Random Forest. Again, Pareto does not show any trend.

[FIGURES 2-3]

Based on this summary information and the results presented in the previous section, we derive the following broad indications for practitioners when addressing issues of income item non-response for the measurement of inequality:

22

1. The use of the sample mean to impute missing items always produces severely downward biased inequality estimates, even when the probability of non-response is disproportionately higher in the middle of the income distribution (close to the mean).

2. With MCAR data, deletion and reweighting are the best options. These two methods result in Gini differences that are not significant for any share of missing incomes. This means that forcing corrections on missing incomes that are distributed randomly with other correction methods may worsen inequality estimates.

3. Reweighting is, on average, the best correction method. It performs better than other methods in the frequently encountered case of MAR data (with MNAR features). However, it does not perform well with MNAR/MID and is outperformed by PMM with MNAR/LEFT and MNAR/RIGHT and by Pareto with MNAR/RIGHT and MNAR/TAILS. Therefore, there is no method that outperforms all others in all cases considered.

4. Corrections based on the Pareto distribution always produce upward biased inequality estimates. While the exact extent of this bias depends on the specific distribution of observed and missing incomes, this is due to the fact that this method corrects (or over corrects) incomes at the top but does not correct for missing incomes in lower parts of the distribution resulting in excess weight given to top incomes. Pareto is the best performing method when the probability of missing incomes is disproportionately higher on the right tail or on both tails of the income distribution.

5. Single and multiple imputation methods, despite their popularity, they never result in the best methods under any data correction pattern and their Gini difference is always significant across all missing incomes shares. They seem to perform better with with MNAR/MID and MNAR/LEFT but these two cases are not particularity important when inequality is measured with income.

6. The use of ML algorithms reduces the prediction error, but it will also limit the variability of predicted incomes which ultimately reinforces downward biases. We observed this phenomenon across missing data patterns with the exception of MNAR/MID where these methods perform relatively well, particularly tree based methods. ML may be more useful when information to predict missing incomes is abundant, which can lead other prediction methods to overfit the model and induce an upward bias in

inequality. This was not the case with the data used by this paper but it is possible with richer data.

Some factors can help to intuitively explain these results. The Reweighting method is somehow unique, because it first estimates the probability of non response from observed data, and then uses the inverse of this probability to reweight observations. It is the only method we know of that explicitly makes an effort to understand the non-response pattern, which is the essential information needed to correct for missing incomes correlated with income. This method performs relatively better also because it estimates the variance of the full distribution (observed and unobserved values) leveraging within group and between groups income variance.

Deletion or ignoring non-responses perform well under the conditions we should expect, with MCAR and, to a lesser extent, MNAR/LEFT data. Observations in the middle of the distribution have less leverage on summary inequality measures than observations located in the tails (even if the Gini attributes more weight to observations located in the middle of the distribution), and observations located in the left tail have less leverage than observations located in the right tail. Perhaps more striking is the fact that this method does not perform too poorly with MAR, MNAR/RIGHT and MNAR/TAIL data. While this method should not be a primary choice, in these latter cases it performs better than some popular methods such as Single and Multiple Imputation and also ML methods.

It is also clear why correcting with mean income cannot perform well. Mean income is typically located towards the center of the distribution, even if it does not necessarily coincide with the median value. When non-responses are concentrated on the tails, as often the case with income distributions, replacing these responses with means results in the maximum possible error. This error can be reduced if means by population sub-groups are used in place of sample mean but a mean would always reduce variability of incomes.

Methods relying on linear predictions can perform relatively well with rich sets of covariates that increase the initial explanatory power of the model and exploit their ability to selectively reduce models with techniques such as regularization. RMSE scores are generally good when compared with other methods but we showed that these positive outcome comes at a cost in terms of precision of the inequality estimate. Methods like Random forests are, for example, the result of averaging over repeated tests, making results more likely to bear external validity but less likely to be close to the real estimates. Overall, these methods are data-specific and by definition ignore possible patterns of income non-

responses.

PMM performs well when other methods tend to underestimate inequality. Among prediction approaches, PMM is the only method together with reweighting that explicitly takes into consideration the part of the variability that cannot be explained by observable covariates by assigning the entire income of similar individuals rather than predicting the income net of an interpolation error term. This makes the method prone to large out-of-sample prediction errors but relatively precise in estimating inequality. In this respect PMM and Reweighting methods are similar to methods proposed by the cross-survey imputation literature which use the empirical or normal distribution of incomes to generate the error term that is not estimated by the prediction model and add the latter to the estimated error to obtain the full variance of the distribution (**?**).

Based on the findings above, to obtain robust inequality measures, a pragmatic approach consists of using more than one correction method and then produce a range of possible estimates. In our case, using Pareto, Reweigting and PMM to define the range would have produced intervals containing the real level of inequality in all simulated missing patterns. With time constraint and the availability of information on non-response rates by geographical area, reweighting should be the method of choice for MAR data with MNAR features.

It is important to caution from taking these results as definitive. This paper is based on one specific data set that was selected for its extreme income inequality, a useful feature for the type of exercise we proposed. Some information on the external validity of our results can be gathered from the RMSE values reported in all tables but, overall, the question of external validity remains unexplored and future research should try to replicate this exercise with other empirical or parametric data to validate results. It is possible, for example, that when initial inequality is much lower, measurement biases may appear less sizable or relevant. In other cases, empirical distributions which (in the absence of item non-responses) would be extremely left or right skewed would probably show different results from distributions that are more uniformly distributed, given that the data corruption patterns proposed focus on different parts of the distribution.

There are pros and cons in replicating our paper with empirical or parametric distributions. With empirical data, income non-responses or misreportings are almost always inevitable, whereas data with very low item non-response rates, such as the one we use, are few and may have different item non-response rates and patterns. In other words, findings from any replication of our paper with other data would still remain limited in terms of

25

external validity, although similar results to ours would provide some degree of validation. Using an income distribution extracted from a parametric function is implicitly biased, as it does not represent any real data, and it is mainly driven by the selection of functions and parameters made by the researcher, which is ultimately a subjective choice. In sum, the best option for future research is probably to replicate our experiment with multiple data sets. This would allow the imposition of different data corruption patterns, as we did, and the comparison of results across data sets.

References

Alfons, A. and M. Templ (2012). Estimation of social exclusion indicators from complex surveys: The r package laeken. *KU Leuven, Faculty of Business and Economics Working Paper*.

Argent, J. (2009). Household income: Report on nids wave 1. Technical report, NIDS, University of Cape Town.

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*, 685–725.

Atkinson, A., T. Piketty, and E. Saez (2011). Top incomes in the long run of history. *Journal of Economic Literature 49*, 3–71.

Bertsimas, D., C. Pawlowski, and Y. D. Zhuo (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research 18*(196), 1–39.

Blackburn, M. L. (2007). Estimating wage differentials without logarithms. *Labour Economics 14*(1), 73–98.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Buuren, S. v. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Ceriani, L. and P. Verme (2021). Population changes and the measurement of inequality. *Social Indicators Research*.

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*(2), 1044–1072.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro–level estimation of poverty and inequality. *Econometrica 71*(1), 355–364.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Greenlees, J. S., W. S. Reece, and K. D. Zieschang (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association 77*(378), 251–261.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. New York, NY: Springer.

Hlasny, V. (2020). Nonresponse bias in inequality measurement: Cross-country analysis using luxembourg income study surveys. *Social Science Quarterly 101*(2), 712–731.

Hlasny, V., L. Ceriani, and P. Verme (2021). Bottom incomes and the measurement of poverty and inequality. *Review of Income and Wealth*.

Hlasny, V. and P. Verme (2018). Top incomes and inequality measurement: A comparative analysis of correction methods using the eu silc data. *Econometrics 6*(2), 1–21.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics 15*(3), 651–674.

Hothorn, T. and A. Zeileis (2015). partykit: A modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research 16*(1), 3905–3909.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica 84*(334), 261–289.

Kennickell, A. B. (2017). Look again: Editing and imputation of scf panel data. *Statistical Journal of the IAOS 33*(1), 195–202.

Korinek, A., J. A. Mistiaen, and M. Ravallion (2006). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality 4*(1), 33–55.

Lillard, L., J. P. 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), 489–506.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association 83*(404), 1198–1202.

Munoz, E. and S. Morelli (2021a). kmr: A command to correct survey weights for unit nonresponse using groups' response rates. *The Stata Journal 21*(1), 206–219.

Munoz, E. and S. Morelli (2021b). kmr: A command to correct survey weights for unit nonresponse using groups' response rates. *The Stata Journal 21*(1), 206–219.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology 27*(1), 85–96.

Riphahn, R. T. and O. Serfling (2005). Item non-response on income and wealth questions. *Empirical Economics 30*(2), 521–538.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, Volume 1, pp. 20–34. American Statistical Association.

Rubin, D. B. (1983). *Imputing Income in the CPS: Comments on "Measures of Aggregate Labor Cost in the United States"*, pp. 333–344. University of Chicago Press.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics 4*(1), 87–94.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association 91*(434), 473–489.

Schenker, N. and J. M. Taylor (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis 22*(4), 425–446.

Schouten, R. M., P. Lugtig, and G. Vink (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation 88*(15), 2909–2930.

Solt, F. (2009). Standardizing the world income inequality database. *Social Science Quarterly 90*(2), 231–242.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science 25*(1), 1–21.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B 58*(1), 267–88.

Vandewalle, B., J. Beirlant, A. Christmann, and M. Hubert (2007). A robust estimator for the tail index of pareto-type distributions. *Computational Statistics & Data Analysis 51*(12), 6252–6268.

Varian, H. R. (2014, May). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

5 Tables

Table 1: Statistics from Complete Original Data

N	Mean	Gini	MLD	Atkinson	Share p20	Share p80	p80p20
7199	8057.23	53.46	0.5210	0.6892	0.0350	0.5900	16.87

Source: NIDS 5. N shows the sample size, MLD stands for mean logarithmic deviation, and p20, p80 and p80p20 respectively denote the 20th and 80th percentile, and their ratio.

Table 2: Gini Difference and RMSE (MCAR)

Gini diff.

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	0.02	-0.72*	-0.30*	-0.23*	-0.02*
6839	5	-0.01	-1.25*	-0.55*	-0.43*	-0.08*
6695	7	-0.01	-1.78*	-0.76*	-0.60*	-0.11*
6479	10	-0.02	-2.64*	-1.10*	-0.88*	-0.12*
6119	15	-0.01	-4.11*	-1.63*	-1.31*	-0.20*
5759	20	0.03	-5.70*	-2.15*	-1.74*	-0.17*
5039	30	-0.02	-9.43*	-3.36*	-2.79*	-0.34*
4319	40	-0.05	-13.72*	-4.52*	-3.84*	-0.44*
3600	50	0.00	-18.57*	-5.69*	-4.94*	-0.54*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	2.08*	0.09	-0.21*	-0.19*	-0.36*
6839	5	2.05*	0.01	-0.37*	-0.45*	-0.57*
6695	7	2.02*	-0.02	-0.55*	-0.77*	-0.82*
6479	10	2.16*	0.03	-0.81*	-1.15*	-1.11*
6119	15	2.12*	0.04	-1.21*	-1.22*	-1.68*
5759	20	2.19*	-0.02	-1.68*	-2.34*	-2.34*
5039	30	2.37*	0.05	-2.71*	-4.20*	-3.92*
4319	40	2.67*	-0.04	-3.79*	-5.06*	-5.46*
3600	50	2.92*	0.00	-4.95*	-7.42*	-7.23*

RMSE

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	n.a.	1519*	1237*	1219*	1711*
6839	5	n.a.	2040*	1667*	1656*	2234*
6695	7	n.a.	2455*	1995*	1996*	2682*
6479	10	n.a.	2988*	2454*	2459*	3292*
6119	15	n.a.	3630*	2996*	3001*	4050*
5759	20	n.a.	4178*	3459*	3465*	4635*
5039	30	n.a.	5150*	4270*	4290*	5663*
4319	40	n.a.	5921*	4918*	4947*	6572*
3600	50	n.a.	6598*	5500*	5536*	7278*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	7118*	n.a.	1205*	1052*	1156*
6839	5	7300*	n.a.	1594*	2209*	1610*
6695	7	7349*	n.a.	1960*	2314*	1886*
6479	10	7453*	n.a.	2399*	2596*	2296*
6119	15	7366*	n.a.	2954*	3051*	2762*
5759	20	7509*	n.a.	3424*	3960*	3276*
5039	30	8033*	n.a.	4220*	4945*	4133*
4319	40	9062*	n.a.	4854*	5393*	4836*
3600	50	9507*	n.a.	5446*	5847*	5384*

Note: NIDS 5. N shows the sample size, and PMM stands for Predictive Means Matching. The star (*) indicates that the number is statistically different from the baseline estimate shown in Table 1, at 01.

Table 3: Gini Difference and RMSE (MAR)

Gini Diff.

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	-0.72*	-1.68*	-0.53*	-0.86*	-0.41*
6839	5	-1.00*	-2.57*	-0.80*	-1.26*	-0.50*
6695	7	-1.33*	-3.50*	-1.11*	-1.68*	-0.74*
6479	10	-1.60*	-4.66*	-1.43*	-2.13*	-0.84*
6119	15	-2.07*	-6.66*	-1.94*	-2.81*	-1.03*
5759	20	-2.45*	-8.64*	-2.47*	-3.45*	-1.18*
5039	30	-3.32*	-13.05*	-3.50*	-4.61*	-1.67*
4319	40	-3.86*	-17.58*	-4.31*	-5.46*	-1.94*
3600	50	-4.44*	-22.65*	-5.20*	-6.26*	-2.16*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	4.35*	-0.14*	-0.87*	-1.21*	-1.23*
6839	5	4.08*	-0.16*	-1.35*	-1.88*	-1.77*
6695	7	1.80*	-0.26*	-1.71*	-2.31*	-2.46*
6479	10	2.54*	-0.26*	-2.36*	-2.39*	-3.05*
6119	15	4.19*	-0.38*	-3.02*	-2.83*	-3.78*
5759	20	3.92*	-0.48*	-3.73*	-3.16*	-4.75*
5039	30	2.31*	-1.15*	-5.05*	-4.81*	-7.01*
4319	40	4.10*	-1.30*	-6.23*	-8.40*	-8.33*
3600	50	0.85*	-1.62*	-7.49*	-10.75*	-10.46*

RMSE

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	n.a.	3224*	2772*	2661*	3335*
6839	5	n.a.	3879*	3352*	3216*	4154*
6695	7	n.a.	4405*	3809*	3652*	4610*
6479	10	n.a.	4927*	4236*	4064*	5202*
6119	15	n.a.	5642*	4866*	4672*	5951*
5759	20	n.a.	6198*	5332*	5142*	6585*
5039	30	n.a.	7133*	6165*	5970*	7447*
4319	40	n.a.	7730*	6736*	6533*	8069*
3600	50	n.a.	8217*	7148*	6994*	8571*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	10208*	n.a.	2658*	2934*	2750*
6839	5	9233*	n.a.	3214*	3967*	3154*
6695	7	4809*	n.a.	3592*	4116*	3513*
6479	10	6074*	n.a.	4118*	4309*	3825*
6119	15	8713*	n.a.	4675*	4639*	4236*
5759	20	8075*	n.a.	5137*	4994*	4663*
5039	30	4808*	n.a.	5922*	5623*	5435*
4319	40	8130*	n.a.	6447*	5826*	5827*
3600	50	3167*	n.a.	6879*	6598*	6304*

Note: NIDS 5. N shows the sample size, P stands for Parametric, PMM stands for
Predictive Means Matching and RF for Random Forest. The star (*) indicates that the
number is statistically different from the baseline estimate shown in Table 1, at 0.001.

Table 4: Gini Difference and RMSE (MNAR/MID)

Gini diff.						
N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	0.30*	-0.60*	-0.18*	-0.03*	-0.19*
6839	5	0.49*	-1.03*	-0.29*	-0.06*	-0.29*
6695	7	0.68*	-1.49*	-0.41*	-0.09*	-0.48*
6479	10	0.96*	-2.23*	-0.60*	-0.14*	-0.62*
6119	15	1.42*	-3.58*	-0.90*	-0.23*	-0.97*
5759	20	1.87*	-5.11*	-1.18*	-0.31*	-1.28*
5039	30	2.77*	-8.61*	-1.68*	-0.45*	-1.71*
4319	40	3.64*	-12.83*	-2.19*	-0.61*	-2.67*
3600	50	4.47*	-17.60*	-2.25*	-0.70*	-3.31*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	4.26*	0.32*	-0.02*	-0.13	-0.22
6839	5	5.99*	0.51*	-0.04*	-0.20	-0.34
6695	7	7.66*	0.70*	-0.04*	-0.31*	-0.45
6479	10	6.25*	0.97*	-0.06*	-0.41*	-0.62*
6119	15	5.35*	1.44*	-0.09*	-0.51*	-0.89*
5759	20	3.36*	1.90*	-0.12*	-0.79*	-1.31*
5039	30	4.16*	2.84*	-0.23*	-1.13*	-2.00*
4319	40	7.34*	3.82*	-0.35*	-1.85*	-2.65*
3600	50	5.65*	4.80*	-0.58*	-1.71*	-3.29*
RMSE						
N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	n.a.	862*	863*	715*	1397*
6839	5	n.a.	1112*	1124*	929*	1816*
6695	7	n.a.	1325*	1353*	1121*	2290*
6479	10	n.a.	1595*	1640*	1364*	2720*
6119	15	n.a.	1991*	2056*	1725*	3415*
5759	20	n.a.	2349*	2440*	2073*	3973*
5039	30	n.a.	3012*	3173*	2757*	5099*
4319	40	n.a.	3671*	3910*	3438*	6174*
3600	50	n.a.	4377*	4711*	4214*	7030*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	10749*	n.a.	701*	826*	762*
6839	5	15369*	n.a.	917*	1019*	946*
6695	7	21458*	n.a.	1084*	1230*	1140*
6479	10	17302*	n.a.	1307*	1554*	1359*
6119	15	15072*	n.a.	1614*	2069*	1661*
5759	20	9783*	n.a.	1903*	2379*	1950*
5039	30	12445*	n.a.	2407*	3196*	2509*
4319	40	25797*	n.a.	2909*	3612*	3081*
3600	50	19573*	n.a.	3400*	4474*	3743*

Note: NIDS 5. N shows the sample size, and PMM stands for Predictive Means Matching. The star (*) indicates that the number is statistically different from the baseline estimate shown in Table 1, at 01.

Table 5: Gini Difference and RMSE (MNAR/LEFT)

Gini Diff.

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	0.03*	-0.50*	-0.21*	-0.10*	-0.01
6839	5	0.05*	-0.89*	-0.37*	-0.18*	-0.04
6695	7	0.05*	-1.32*	-0.55*	-0.28*	-0.06*
6479	10	0.06*	-2.00*	-0.83*	-0.43*	-0.08*
6119	15	0.09*	-3.31*	-1.29*	-0.69*	-0.11*
5759	20	0.13*	-4.79*	-1.77*	-0.96*	-0.10*
5039	30	0.20*	-8.35*	-2.81*	-1.57*	-0.14*
4319	40	0.27*	-12.69*	-3.82*	-2.21*	-0.08*
3600	50	0.30*	-17.87*	-4.81*	-2.88*	-0.10*

N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	3.61*	0.07*	-0.10*	-0.15	-0.22
6839	5	4.91*	0.11*	-0.17*	-0.29*	-0.36*
6695	7	6.41*	0.14*	-0.24*	-0.45*	-0.52*
6479	10	8.31*	0.20*	-0.36*	-0.67*	-0.80*
6119	15	3.40*	0.31*	-0.57*	-1.01*	-1.22*
5759	20	5.21*	0.45*	-0.79*	-1.50*	-1.69*
5039	30	3.70*	0.76*	-1.33*	-2.19*	-2.82*
4319	40	6.74*	1.11*	-1.96*	-3.00*	-3.94*
3600	50	5.30*	1.46*	-2.69*	-3.62*	-5.12*

RMSE

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	n.a.	697*	575*	462*	1002*
6839	5	n.a.	925*	778*	628*	1284*
6695	7	n.a.	1120*	939*	763*	1569*
6479	10	n.a.	1379*	1146*	945*	1914*
6119	15	n.a.	1776*	1499*	1268*	2513*
5759	20	n.a.	2145*	1801*	1554*	3003*
5039	30	n.a.	2880*	2400*	2144*	3973*
4319	40	n.a.	3669*	3021*	2804*	4909*
3600	50	n.a.	4566*	3689*	3525*	6068*

N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	9197*	n.a.	459*	569*	574*
6839	5	12440*	n.a.	609*	762*	718*
6695	7	16761*	n.a.	736*	906*	853*
6479	10	24340*	n.a.	902*	1136*	1029*
6119	15	8926*	n.a.	1148*	1453*	1281*
5759	20	14510*	n.a.	1376*	1814*	1569*
5039	30	10521*	n.a.	1805*	2483*	2056*
4319	40	21474*	n.a.	2246*	3107*	2640*
3600	50	17294*	n.a.	2711*	3796*	3341*

Note: NIDS 5. N shows the sample size, and PMM stands for Predictive Means Matching. The star (*) indicates that the number is statistically different from the baseline estimate shown in Table 1, at 0.1.

Table 6: Gini Difference and RMSE (MNAR/RIGHT)

Gini Diff.

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	-4.73*	-5.73*	-4.59*	-4.96*	-4.49*
6839	5	-5.43*	-6.90*	-5.33*	-5.79*	-5.27*
6695	7	-6.09*	-8.02*	-6.02*	-6.55*	-5.85*
6479	10	-6.70*	-9.32*	-6.72*	-7.33*	-6.38*
6119	15	-7.51*	-11.36*	-7.72*	-8.42*	-7.14*
5759	20	-8.15*	-13.34*	-9.09*	-9.39*	-7.76*
5039	30	-9.12*	-17.32*	-10.62*	-11.06*	-8.54*
4319	40	-9.88*	-21.69*	-11.82	-12.52*	-9.06*
3600	50	-10.50*	-26.50*	-13.31*	-13.92*	-9.44*

N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	2.64*	-3.39*	-5.03*	-5.55*	-5.16*
6839	5	6.56*	-4.06*	-5.99*	-6.34*	-6.14*
6695	7	3.68*	-4.79*	-6.77*	-6.79*	-6.86*
6479	10	5.80*	-5.48*	-7.69*	-8.10*	-8.09*
6119	15	2.82*	-6.40*	-8.79*	-9.55*	-9.63*
5759	20	4.21*	-7.12*	-9.78*	-11.42*	-10.95*
5039	30	0.75*	-8.23*	-11.50*	-13.52*	-13.12*
4319	40	3.88*	-9.23*	-13.03*	-14.56*	-15.29*
3600	50	3.15*	-9.89*	-14.57*	-17.56*	-17.35*

RMSE

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	n.a.	6865*	5724*	6013*	6021*
6839	5	n.a.	7215*	5988*	6286*	6340*
6695	7	n.a.	7460*	6171*	6471*	6546*
6479	10	n.a.	7721*	6366*	6666*	6787*
6119	15	n.a.	8046*	6616*	6906*	7087*
5759	20	n.a.	8287*	6801*	7081*	7289*
5039	30	n.a.	8652*	7087*	7335*	7653*
4319	40	n.a.	8931*	7317*	7522*	7927*
3600	50	n.a.	9152*	7503*	7667*	8142*

N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	3819*	n.a.	6118*	6024*	5728*
6839	5	8759*	n.a.	6407*	6257*	6044*
6695	7	4668*	n.a.	6608*	6419*	6260*
6479	10	7195*	n.a.	6824*	6726*	6529*
6119	15	3309*	n.a.	7082*	7021*	6809*
5759	20	4705*	n.a.	7291*	7347*	7036*
5039	30	1315*	n.a.	7596*	7674*	7407*
4319	40	3304*	n.a.	7830*	7812*	7724*
3600	50	2660*	n.a.	8038*	8096*	7949*

Note: NIDS 5. N shows the sample size, and PMM stands for Predictive Means Matching. The star (*) indicates that the number is statistically different from the baseline estimate shown in Table 1, at 01.

Table 7: Gini Difference and RMSE (MNAR/TAIL)

Gini Diff.

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	-4.08*	-4.93*	-3.98*	-4.24*	-3.92*
6839	5	-4.82*	-6.07*	-4.75*	-5.07*	-4.64*
6695	7	-5.31*	-6.94*	-5.30*	-5.64*	-5.11*
6479	10	-5.96*	-8.20*	-6.04*	-6.41*	-5.75*
6119	15	-6.83*	-10.11*	-7.08*	-7.48*	-6.53*
5759	20	-7.53*	-11.92*	-7.96*	-8.36*	-7.17*
5039	30	-8.63*	-15.55*	-9.60*	-9.94*	-8.19*
4319	40	-9.65*	-19.53*	-11.27*	-11.51*	-9.12*
3600	50	-10.50*	-23.85*	-12.73*	-12.88*	-9.88*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	3.95*	-3.36*	-4.28*	-4.52*	-4.32*
6839	5	2.13*	-4.22*	-5.15*	-5.64*	-5.34*
6695	7	4.43*	-4.78*	-5.82*	-6.44*	-6.21*
6479	10	6.06*	-5.54*	-6.60*	-6.94*	-7.19*
6119	15	3.65*	-6.55*	-7.78*	-8.12*	-8.31*
5759	20	5.14*	-7.33*	-8.64*	-9.47*	-9.71*
5039	30	3.41*	-8.63*	-10.30*	-11.81*	-12.29*
4319	40	1.69*	-9.75*	-11.85*	-14.99*	-14.83*
3600	50	3.56*	-10.67*	-13.41*	-16.38*	-16.97*

RMSE

N	% Miss.	Deletion	Mean	SingleImp.	Mult.Imp.	PMM
6983	3	n.a.	6529*	5484*	5755*	5739*
6839	5	n.a.	6879*	5733*	6015*	6063*
6695	7	n.a.	7110*	5908*	6195*	6267*
6479	10	n.a.	7365*	6100*	6385*	6475*
6119	15	n.a.	7673*	6329*	6606*	6755*
5759	20	n.a.	7906*	6510*	6772*	6978*
5039	30	n.a.	8249*	6779*	7012*	7306*
4319	40	n.a.	8522*	6998*	7194*	7572*
3600	50	n.a.	8737*	7170*	7323*	7795*
N	% Miss.	Pareto	Reweight	LASSO	Tree	RF
6983	3	6049*	n.a.	5835*	5655*	5476*
6839	5	3371*	n.a.	6124*	6036*	5734*
6695	7	5764*	n.a.	6309*	6232*	5917*
6479	10	7999*	n.a.	6523*	6175*	6221*
6119	15	4505*	n.a.	6767*	6636*	6587*
5759	20	6144*	n.a.	6948*	6901*	6841*
5039	30	3875*	n.a.	7235*	7183*	7103*
4319	40	2067*	n.a.	7448*	7636*	7312*
3600	50	3357*	n.a.	7629*	7765*	7521*

Note: NIDS 5. N shows the sample size, and PMM stands for Predictive Means Matching. The star (*) indicates that the number is statistically different from the baseline estimate shown in Table 1, at 01.

Table 8: Summary Results

Methods	MCAR	MAR	MNAR-M	MNAR-L	MNAR-R	MNAR-T
Deletion						
Gini	0.00	-2.31	1.84	0.13	-7.57	-7.03
Mean						
RMSE	3831	5706	2255	2129	8037	7663
Gini	-6.43	-9.00	-5.90	-5.74	-1.34	-11.90
Sing. Par.						
RMSE	3166	4935	2363	1761	6619	6334
Gini	-2 .23	-2.37	-1.10	-1.83	-8.27	-7.63
Mult. Par.						
RMSE	3174	4767	2037	1566	6883	6584
Gini	-1.86	-3.17	-0.29	-1.03	-8.23	-7.95
PMM						
RMSE	4235	5991	3768	2915	7088	6772
Gini	-0.22	-1.16	1.28	-0.01	-7.10	-6.70
Pareto						
RMSE	7855	7024	16394	15051	4415	4793
Gini	2.28	3.12	5.56	5.29	3.72	3.78
Reweighting						
Gini	0.00	-0.64	1.92	0.51	-6.51	-6.76
LASSO						
RMSE	3117	4738	1805	1332	7088	6758
Gini	-1.81	-3.53	-0.17	-0.91	-9.24	-8.20
Tree						
RMSE	3486	4778	2262	1781	7042	6691
Gini	-2.53	-4.19	-0.78	-1.43	-10.38	- 9.37
Random Forest						
RMSE	3040	4414	1906	1562	6831	6524
Gini	-2.61	-4.76	-1.31	-1.85	-10.29	-9.46

Source: NIDS 5. RMSE stands for Root Mean Squared Error. Sing. Par. stands for single parametric, Mult. Par. stands for Multiple parametric, PMM stands for Predictive Means Matching.
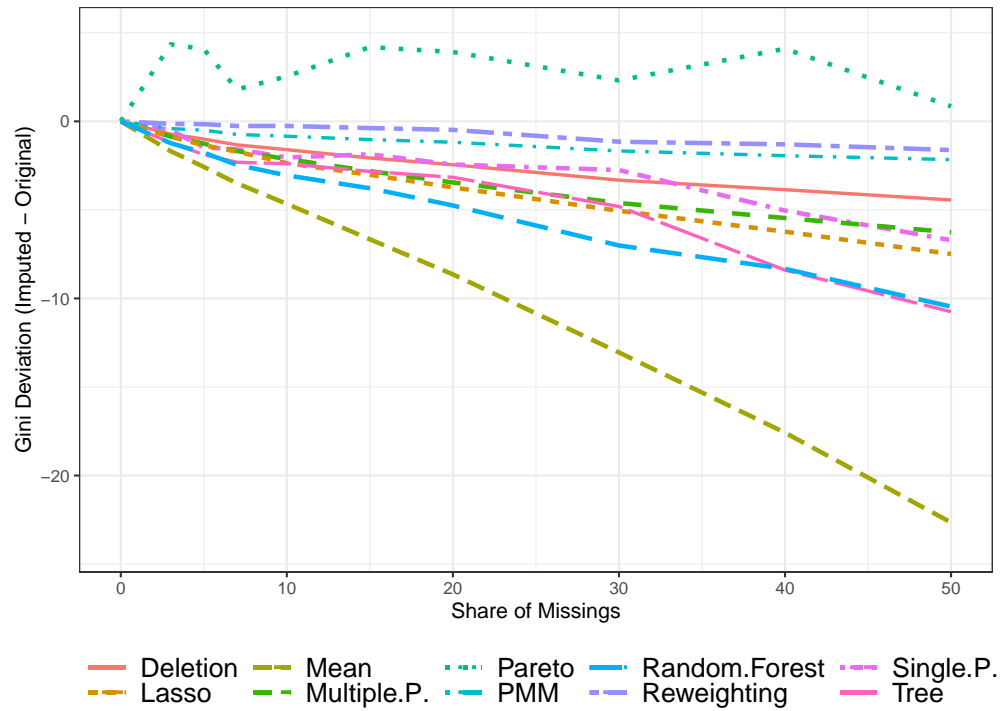
6 Figures

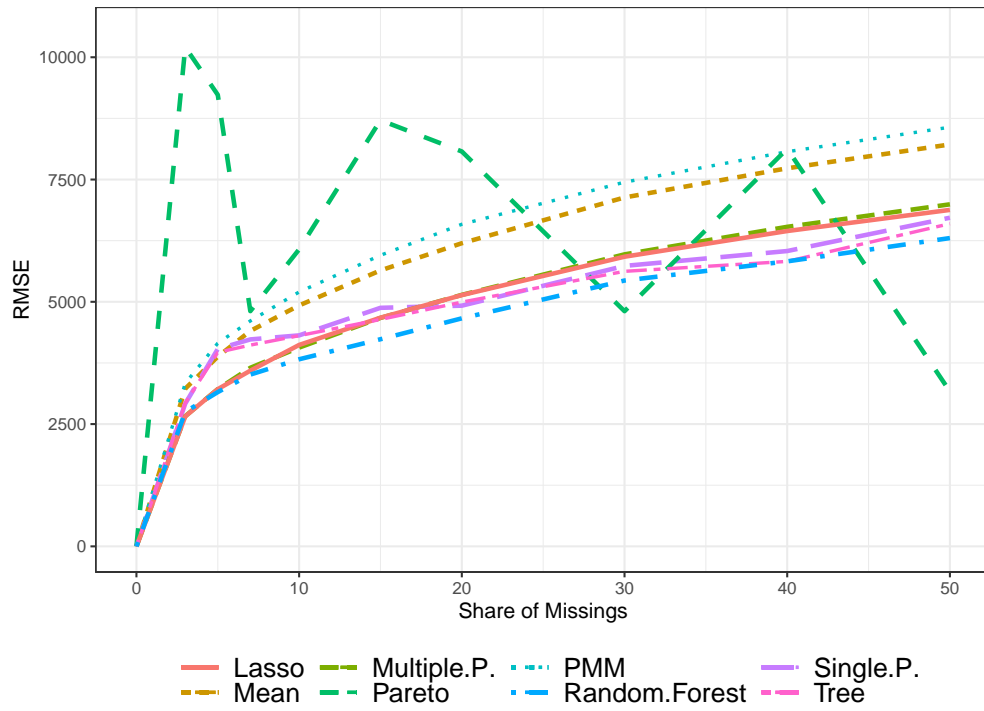Figure 1: Conditional missing probabilities under MNAR
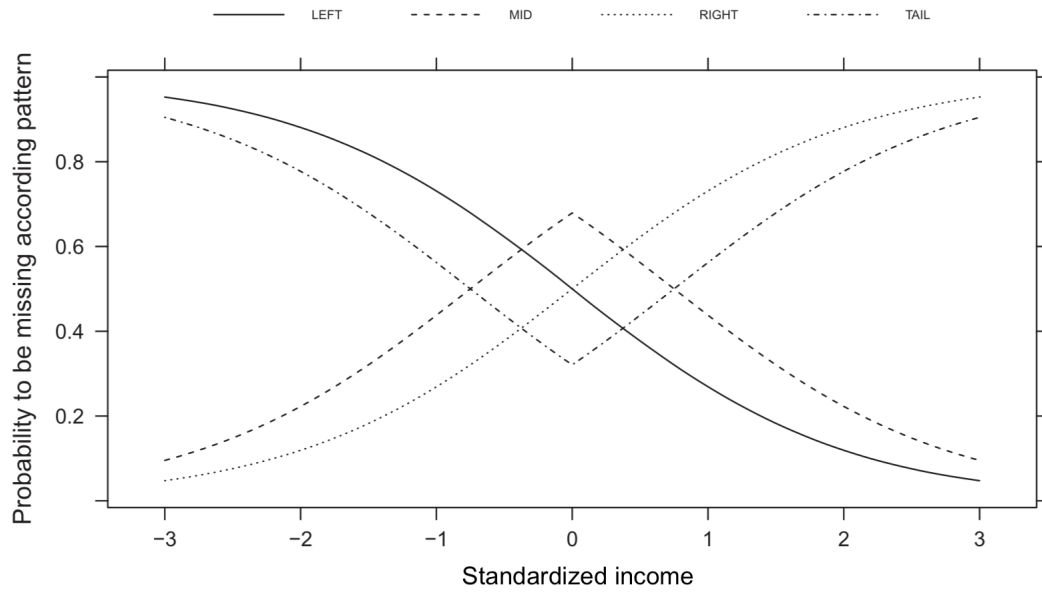
Source: Figure from Schouten et al. (2018).

Figure 2: Gini Deviation by Shares

Source: NIDS 5. P. stands for Parametric, PMM stands for Predictive Means Matching.

Figure 3: RMSE by Shares

Legend: Lasso, Mean, Multiple.P., Pareto, PMM, Random.Forest, Single.P., Tree

Source: NIDS 5. P. stands for Parametric, PMM stands for Predictive Means Matching.

A Appendix

A.1 LASSO MAR deletion

Table A1 shows the coefficients for the LASSO regularization, where the dependent is a binary variable that takes 1 if the observation was originally missing and 0 if it was observed. We use the deviance of the binomial model for the tuning of the Lasso. The lambda selected is equal to 0.0016. (log of lambda = -6.437752).
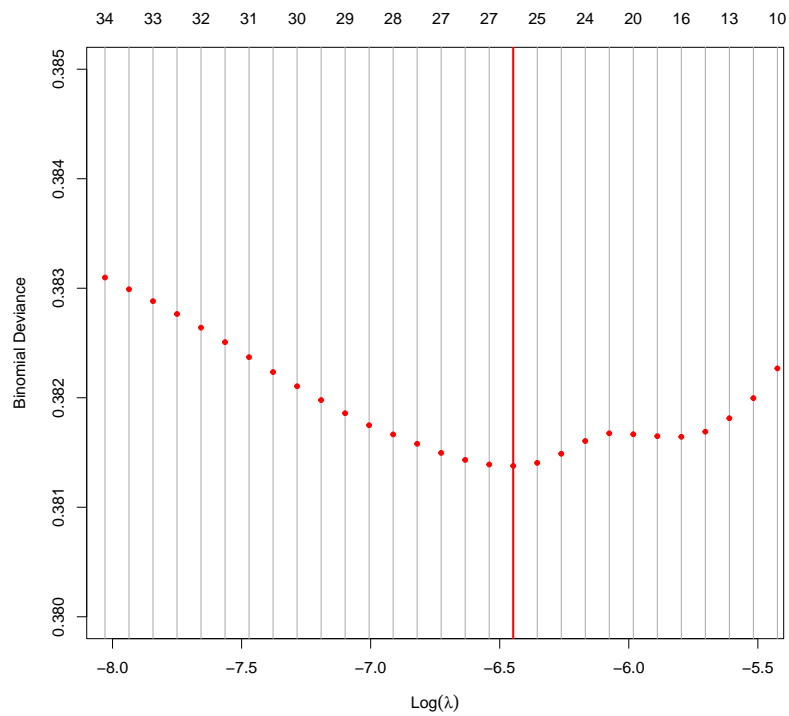
Figure A1: Lasso tuning

Table A1: Lasso Estimates

	Coefficient
Male	0.3372
Asian or Indian	0.4555
White	0.5427
Trade Union	-0.3427
Schooling	0.0546
Cert (nomat)	-0.1010
Dipl (nomat)	-0.1277
Cer (mat)	-0.1926
Dip (mat)	0.0364
Bachelors	0.0753
Honours	-0.2936
Married	-0.0156
Homerooms	0.1404
Homerooms squared	-0.0031
Int. Month (May)5	-0.0901
Int. Month (June)6	0.0000
Int. Month (July)7	-0.0632
Int. Month (October)10	0.2932
Int. Month (November)11	0.2728
Int. Month (December)12	1.1653
Province: 2	0.1554
Province: 3	-0.0371
Province: 4	0.1854
Province: 6	0.2383
Province: 7	0.0186
Province: 8	-0.0023
Province: 9	0.2176
Geo Type: 3	-0.8345

Note: Source: NIDS 5.

Table A2: Packages

Method	R Package	Reference
Data corruption	ampute	Schouten et al., 2018
Imputation of the mean	mice	Buuren and Groothuis-Oudshoorn, 2011
Multiple Imputation	mice	Buuren and Groothuis-Oudshoorn, 2011
PMM	mice	Buuren and Groothuis-Oudshoorn, 2011
Parametric distribution	laeken	Alfons and Templ, 2012
Sample reweighting	krm (Stata)	Munoz and Morelli, 2021b
LASSO	glmnet	Friedman et al., 2010
Tree	partykit	Hothorn and Zeileis, 2015
Random forest	partykit	Hothorn and Zeileis, 2015

Note: All packages are implemented in R with the exception of "kmr", that is implemented in Stata.