Survey Research Methods (2023)
Vol. 17, No. 2, pp. 159-175
doi:10.18148/srm/2023.v17i2.7949
European Survey Research Association

Income Imputation in Longitudinal Surveys: A Within-Individual Panel-Regression Approach

Oliver Lipps and Ursina Kuhn
FORS

Unlike for cross-sectional data, there is only little research on income imputation for long-running panel surveys. In this contribution, we test different longitudinal imputation methods: Little and Su (LS) method, iterative regression with lagged income, and a new imputation method ("mean and within" imputation). The LS method is a univariate approach based on individual mean income over time and is the current best practice for long-running panel data. Iterative regression, which is also frequently used for longitudinal data, has the advantage of using additional information in the wave with missing data. The mean and within approach is based on the individual mean income like the LS method, and adds a component for within variation using iterative regression similar to iterative regression. We evaluate the different imputation methods including complete case analysis using employment income from the Swiss Household Panel from 2000–2021. The nonresponse mechanisms used for the evaluation is based on an external data source containing both registry information and survey questions on income, allowing to detect a not missing at random mechanism. We use performance criteria proposed in previous evaluations of longitudinal imputation methods and add the performance in an application example of regression analysis. Our results confirm the good performance of LS for cross-sectional analysis, but more biased estimates for longitudinal analysis, such as income mobility. The mean and within approach performed best for longitudinal criteria. For multivariate regression, imputation does not improve the estimates.

Keywords: longitudinal imputation; row-and-column method; Little and Su; panel data; nonresponse mechanism; Swiss Household Panel

1 Introduction

The importance to impute missing income values is well established (e.g., Champney & Bell, 1982; Durrant, 2009; Little, 1988, 1992). Many surveys provide single or multiple imputed income data to their users (see Aßmann et al., 2017).

Imputation in the longitudinal context is more complex, as imputed values do not only have to take cross-sectional multivariate joint distributions into account but must also address longitudinal aspects such as changes over time between and within individuals. So far, there is no standard for the imputation of longitudinal data. Some surveys, such as the Survey of Health, Ageing and Retirement in Europe (SHARE), the Household Finance and Consumption survey, or the Statistics of Income and Living Conditions in Europe (EU-SILC) pursue a cross-sectional approach (see De Luca et al., 2015, for SHARE) that ignores the longitudinal data structure. In contrast, most household panels that are part of the Cross National Equivalent File (CNEF; see Frick et al., 2007) use data from other panel waves of the same individual for the imputation. The most frequently used method is the so called row-and-column method introduced by Little and Su (1989), denoted as "LS" in the following. Two studies, which compared different imputation methods in long-running household panels (Watson and Starick, 2011 for the Household, Income and Labour Dynamics in Australia (HILDA) Survey and Westermeier and Grabka, 2016 for the German Socio-Economic Panel (SOEP)) report good performance of the LS method, in particular for cross-sectional evaluation criteria.

The LS method has an important shortcoming, which may explain the poorer performance for longitudinal analysis, such as underestimating wealth mobility (Westermeier & Grabka, 2016). The approach assumes that variation of income over time *within* individuals is unrelated to the individual's situation in a particular wave. As a univariate approach, the LS imputation does not employ wave specific information of the individual in the wave with missing income, even though they may help predicting the missing value. For example, a higher satisfaction with income or a professional promotion is likely to be associated with an increased income in this specific wave.

Employing the predictive power of covariates is the basis of multivariate imputation. The gold standard for cross-

Contact information: Oliver Lipps, Swiss Centre of Expertise in the Social Sciences (FORS), c/o University of Lausanne, 1015 Lausanne, Switzerland (E-mail: oliver.lipps@fors.unil.ch)

sectional data are iterative imputations (also referred to as chained equations) or joint modelling. These methods are not well suited for panel data, because dynamics over time, non-monotone missing patterns, refreshment samples, as well as non-linear and incomplete predictors (e.g., lagged values) present additional complexity for the imputation process (Spiess et al., 2021).

To evaluate established practices and potentially improving imputation methods for long-term longitudinal data, we test the most frequently used imputation methods in household panels, standard LS, iterative regression (chained equation) and complete case analysis. In addition, we present and test mean and within imputation as an alternative approach. The imputation is based on two multiplicative components: the mean of the individual observed income, and wave-specific deviation from the individual mean income based on a regression model. Compared with univariate approaches (such as LS), considering auxiliary information from the wave with missing data may not only improve accuracy of the imputation, but capture variation within individuals which is, after all, one of the main reasons to use panel data. We expect that particularly longitudinal analyses may benefit from the information from the wave with missing income. Compared to iterative regression approaches, including the individual-specific mean as a component explicitly acknowledges the panel structure. To test the different imputation methods, we build on evaluation criteria proposed in the related literature (Watson & Starick, 2011; Westermeier & Grabka, 2016), using employment income from the Swiss Household Panel (SHP) spanning over 20 waves (2002–2021).

The contribution of this study to the literature is fourfold. First, we propose and test an imputation method, referred to as mean and within approach, that employs the explanatory potential of covariates. Second, we add evidence on the strength and weaknesses of frequently used longitudinal imputation approaches. Third, this study is based on a not missing at random (NMAR) nonresponse mechanism estimated from the CH-SILC survey that linked survey income with registry income. We transfer the mechanism for unit- and item nonresponse to the SHP data. Finally, testing the properties of different imputation methods also provides a basis to reflect on the utility and limits of providing all-purpose imputed data to data users. While ready-to-use imputations are of great value for the scientific community considering resources and knowledge required for a high-quality imputation, for transparency, and for reproducibility, they can give the false impression that the imputed values are suitable for all data applications.

The next section presents the imputation method in more detail. We then discuss the evaluation criteria and compare the performance of the different imputation models before we conclude.

2 Established imputation methods

2.1 Evaluation studies

Imputation of missing income data in panel studies needs to take account of both cross-sectional and longitudinal imputation strategies. Different approaches have been tested and used in the literature. Watson and Starick (2011) compared carryover, LS, longitudinal hot deck, and different longitudinal nearest neighbour regression methods. The performances of the different imputation methods varied strongly among the different income variables tested (wages and salaries, pension and benefits, business income, total income), criteria of evaluation, and whether values were imputed for respondents or non-respondents. The authors conclude that the LS method works well for item nonresponse, and a combination of carry-over (to determine whether income is zero or positive) and LS (to estimate positive income amounts) work well for unit nonresponse. However, it should be noted that the study did not assess the NMAR response mechanism.

Westermeier and Grabka (2016) tested several multiple imputation approaches for wealth data: imputation by chained equations, regression with Heckman correction for sample selection, and the LS method with and without a distinction by age groups. For cross-sectional analysis, such as trend accuracy and inequality accuracy, LS outperformed the chained equation approach for all assets (home market value, financial assets, consumer credits) and both MAR and MNAR nonresponse mechanisms analysed. The advantage of the LS method was particularly strong under the assumption that data was not missing at random (NMAR). The weakness of the LS was the underestimation of wealth mobility, where the authors recommend an iterative regression approach instead.

Spiess et al. (2021) evaluated multiple regression imputation, LS and last value carried forward for growth curves of juvenile delinquency data. While the last value carried forward approach did not work well, the LS method yields similar means and growth factors as multiple imputation. However, the LS methods underestimated the variances and covariance's of the latent growth factors compared with multiple regression imputation. This study did not assess within-individual dynamics, though.

Based on these reviews and current practices for large panel surveys, we will test the LS method and the iterative regression (chained equation) approach method for the evaluation. Both performed well in these three evaluation studies and are widely used in practice. Furthermore, the mean and within approach builds on the strength of both methods, the LS method for the mean effect and the iterated regression for the within effect. Finally, we compare the imputation methods with complete-case analysis (listwise deletion), as this is the most frequently used approach to analyse data with

missing observations.

2.2 Little and Su

The Little and Su (1989) imputation uses the so-called row-and-column effects. The column effect is calculated for each wave t:

$$c_t = \frac{\overline{y}_t}{\overline{\overline{y}}} \quad \text{where} \quad \overline{\overline{y}} = \frac{1}{T}\sum_{t=1}^{T}\overline{y}_t \quad . \tag{1}$$

T is the number of waves and $\overline{y}_t$ is the sample mean of income at wave t from complete cases.

The row effects are computed for each individual i:

$$r_i = \frac{1}{T_i}\cdot\sum_{t=1}^{T_i}\frac{y_{it}}{c_t} \quad , \tag{2}$$

where T_i is the number of waves that individual i reported income and y_{it} is income of individual i in wave t. The respondents are ordered by r_i, and the incomplete case is matched to the reported value y_{jt} from the same wave with the closest r_j. There are different variants in how the donor is selected. Ideally, the same donor is used for all missing cases of an individual. If response patterns are highly diverse and irregular, as in the case of the SHP due to gaps in response patterns or refreshment samples, this might be too restrictive in the sense that such a donor does not exist or the row effect of donor and recipient are too different.

2.3 Iterative regression

Iterative equation approach is a model-based approach using the predictive power of variables that correlate with income. The procedure is to sequentially impute variables with missing observation using appropriate regression models, which include all variables suitable to predict the variable of interest in addition to the variables to impute. There are many different versions of the iterative regression approach, some being specific to multilevel data (Audigier et al., 2018).

2.4 Mean and within regression as longitudinal imputation method

Our proposed mean and within regression has two multiplicative components. The first is—similar to the row effect in the LS method—the mean reported income of the individual in all waves. The aim of the first component is to give the most realistic value in a univariate setting if at least one value is reported. The second component of the imputation method (within component) is the wave-specific deviation from the individual mean income. The aim of this component is to predict individual-specific deviation considering life circumstances of the individual in the wave with missing income. Specifically, we calculate the predicted value based on a pooled OLS[1] regression in the wave with missing data divided by the individual mean of the predicted values in all waves. Formally, the mean and within regression imputation can be described as follows:

$$\text{imputation}_{it} = \overline{y}_i \cdot \text{dev}_{it} \quad \text{where} \quad \overline{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it} \tag{3}$$

and

$$\text{dev}_{it} = \frac{\widehat{y_{it}}|x_{it}}{\frac{1}{T_i}\sum_{t=1}^{T_i}(\widehat{y_{it}}|x_{it})} \quad , \tag{4}$$

with y_{it} denoting the reported income of a person i at wave t, T_i the number of observed income values (equation 3) or predicted (equation 4) income values of person i and $\widehat{y_{it}}|x_{it}$ the predicted value of y_{it} using an OLS regression model and predictors x.

In case the chosen approach underestimates variation due to regression to the mean, a random component may be added to the imputed value. However, since we do not directly use the predicted value from the regression approach, underestimation of variability does not necessarily occur. Therefore, we will test the between- and the within- variation as part of the evaluation criteria.

Whereas the LS method uses the wave-specific deviation of *another* individual in the panel (matching), we use a panel-regression model to predict this wave-specific deviation from the individual mean income using auxiliary variables. Because regression models are able to explain a part of the variance within individuals over time, we expect that the additional information should improve predictive accuracy (reproduce true values) compared to univariate approaches (such as the LS method). However, the main task of imputation is not predictive accuracy for single individuals, but to produce accurate and efficient estimates for the sample. In addition, the mean and within approach differs to existing extended versions of the LS imputation that look for nearest neighbours within a stratification variable such as age groups. This later approach is only suitable for large datasets as the number of available donors in the imputation reduces with the number of groups. As the second component of the mean and within approach is based on a regression approach, it shares some similarities with the iterative regression approach tested here.

[1] As alternatives, we firstly have also tested fixed effects (FE) models. These models differ in the way the variation within individuals is accounted for. The pooled OLS model does not distinguish between the variation between individuals and the variation within individuals over time while the FE model uses only variation within individuals (Andreß et al., 2013). Secondly, we tested a "row&within" approach, using the row effect (r_i in equation 2). We comment on the findings in the result section.

3 Evaluation criteria

Watson and Starick (2011) proposed six evaluation criteria to compare different imputation methods: prediction, distribution, and estimation accuracy, each evaluated cross-sectionally (income level) and longitudinally (difference between waves) (see Table 1).

Focusing on multiple imputation, Westermeier and Grabka (2016) selected five of the evaluation criteria used by Watson and Starick (2011), the comparison of cross-wave correlations of the true and the imputed values, Kolmogorov-Smirnov distance, income mobility, absolute relative difference in means, and the absolute difference in the coefficient of variation. By adding the Gini coefficient, the mean log deviation and the 99/50 ratio of percentiles, they focused more on inequality indicators. Moreover, they added the relative bias of standard errors.

Building on these previous studies, we selected eight criteria to test the performance of the imputed values: four cross-sectional criteria, three longitudinal criteria and multivariate relationships (Table 1). Multivariate relationship and variation over time, which we consider as central for the use of panel data, were not assessed in previous studies. A few criteria used either by Watson and Starick or Westermeier and Grabka were not retained for this assessment.[2]

For *cross-sectional criteria*, the first tests the accuracy of widely used *descriptive statistics (1)*. These include the mean *(1a)* and inequality measures evaluated in previous studies (*1b–1e*). The *correlation between observed and imputed values (2)* assesses the predictive accuracy. To take account of potential outlier effects, this criterion is assessed on the full sample without the upper 1 or 5 percentile, and the full sample without the lower 1 or 5 percentile, in addition to the full sample. The predictive accuracy is also tested with the *absolute deviation from observed values (3)*. The last cross-sectional criterion is the *Kolmogorov-Smirnov distance (4)* to assess the distributional accuracy. It tests the maximum distance between the imputed and the true distribution functions. Taken together, the cross-sectional criteria encompass estimation accuracy (criterion 1), predictive accuracy (2 and 3) and distributional accuracy (4).

As *longitudinal criteria*, we first compare *cross-wave correlations (5)*, specifically the correlation with the lagged values of the observed and the imputed values. As we did for criterion 2 (correlation between observed and imputed values), we take the possible impact of extreme values into account, by estimating correlation also without the upper and lower 1 or 5 percentile. The second longitudinal criterion is *income mobility (6)*, where we measure stability between income deciles with spearman correlation. To test *variability over time*, we decompose variation into *between-individual (7a)* and *within-individual variation over time (7b)*. These criteria are important as regression approaches might understate variability due to regression to the mean.

While these criteria are univariate, survey data are mostly used to assess relations between variables. Therefore, we test the performance of the imputed income variable in a typical regression model application. *(8)*. We estimate multivariate models and compare the income regression coefficient of observed and imputed data.

4 Methods and Data

4.1 Sample and variables

We test the mean and within regression imputation using data from the Swiss Household Panel (SHP), a probability-based longitudinal survey which started in 1999 (see Tillmann et al., 2016, for details). As an application example, we impute income from employment between the years 2002 and 2021.

For the imputation methods that include multivariate regression (iterative regression, mean and within imputation), we include explanatory variables from three different sources. Firstly, we include socio-demographic information[3] collected in the household grid questionnaire from the household reference person. This information is available even for individuals in the sample, who have not completed the personal interview (i.e., we consider fully and partially reporting households). Secondly, we include information from the household questionnaire[4], also answered by the household reference person. The third group of variables involves job

[2]Following the argument by Westermeier and Grabka (2016), we did not include skewness and kurtosis because this is covered by the Kolmogorov-Smirnov distance. To limit the criteria that evaluate predictive accuracy, we did not assess the relation between the logarithm of imputed and the logarithm of observed values by regression and the Euclidian distance between the imputed and observed data in multi-dimensional space, as prediction at the individual level is not the main aim of imputation. Moreover, we did not retain the relative bias of standard errors (impact of the imputation methods on statistical inference) used by Westermeier and Grabka, and the correlation between two income variables for imputed and true income values used by Watson and Starick (2011).

[3]Number of adults in the household, number of children in the household, urbanity, nationality (Swiss, North-West European countries or USA, or Australia, other nationalities), married (vs. another civil status), living with a partner, years of education, socio-economic status of parents, gender, and employment status (full time, part time, mini job, unemployed, retired).

[4]Self-assessed estimated minimum income to make ends meet, payment arrears, satisfaction with household finances, reception of health insurance subsidies, noisy environment, can go to the dentist if need, having a 3rd pillar (private pension insurance), having a compute, having a car, going to the restaurant, invite friends at least once a month, at least one week of holiday per year, number of rooms of the accommodation, house ownership, saving behaviour (household can save, household spends what it earns, eat savings, gets into debt), how well household finances are evaluated (0–10 scale).

Table 1

Overview of evaluation criteria

	Watson and Starick (2011)	Westermeier and Grabka (2016)
Cross sectional		
1 Descriptive statistics		
a. First moment (mean)	✓	✓
b. Coefficient of variation (CV)	✓	✓
c. Gini coefficient	-	✓
d. Mean log deviation	-	✓
e. 99/50 ratio of percentiles	-	✓
2 Correlation between observed and imputed values	✓	-
3 Absolute deviation from observed values	✓	✓
4 Distribution: Distance between the empirical distribution functions for both the imputed and the true values(Kolmogorov-Smirnov)	✓	✓
Longitudinal		
5 Comparison of cross-wave correlations of the true and the imputed values	✓	✓
6 Income mobility (in deciles) in the imputed data compared with the one with observed values	✓	✓
7 Variability over time		
a. Within individuals	-	-
b. Between individuals	-	-
Multivariate		
8 Parameter accuracy of the income regression coefficient in typical regression applications	-	-

related information[5] from the individual questionnaire. This means, in the case of partial-unit nonresponse (no personal interview), only the first two sets of variables (household grid and household questionnaire) are reported. In the case of item nonresponse, all three sets of variables (household grid, household questionnaire, individual questionnaire) are reported. In addition, we include the survey year in the imputation model.

Our analysis sample comprises all individuals, who reported an income from employment and where the household questionnaire was completed (complete cases). We deleted cases with missing income before starting the simulation (partial unit nonresponse=23%, item nonresponse=3%). The sample with reported income amounts to 102,891 observations from 20,471 individuals.

4.2 Nonresponse mechanism

To test the imputations, we set some of the observed income values artificially to missing, so that we can compare the imputed values to observed (true) values. This requires decisions on how many, and which, observations should be set to missing (nonresponse mechanism). Regarding the former point, longitudinal imputation requires at least one observed value per individual (Frick & Grabka, 2014; Watson & Starick, 2011). In our implementation, we kept a random reported value of each individual valid. Of the remaining observations, we simulate missing values separately for partial unit nonresponse and for item nonresponse. In the following, we explain how we did this.

The nonresponse mechanism is central for such simulations. In general, nonresponse is not completely missing at random (MCAR) but is related to other variables. If the probability of response can be explained by observed variables, and thus does not depend on unobserved variables after conditioning on observed values, data are said to be missing at

[5]Treiman score of main job, change of job in the previous year, change of employer in the previous year, change of job and employer in the previous year, supervision task.

random (MAR). If the nonresponse mechanism depends in addition on unobserved values, data are said to be not missing at random (NMAR). In particular, this is the case if nonresponse depends on the income value itself. As the true income is unknown, NMAR mechanisms cannot be detected or simulated directly based on the data used for the analysis.

To approach this problem, we analyse the amount of nonresponse and the nonresponse mechanism using an external data source, which contains income information even for non-respondents, and apply the identified mechanism to the (complete) SHP data. We use the Swiss part of the Statistics on Income and Living Conditions (CH-SILC) survey from 2016 for this purpose, because this survey has been matched to the income registry from the social security system and is highly comparable to the SHP data.[6] The SILC-data contain information on employment income during a calendar year from both the survey and from the registry. Therefore, we have information on employment income, item nonresponse and partial unit nonresponse for each person in the sample. For the purpose of this paper, this linkage required no additional consent. The reliance on an external data source avoids using the same data for the imputation and the nonresponse mechanism.

For both nonresponse mechanisms, we estimated separate models for partial unit nonresponse and for item nonresponse using logistic regressions in the SILC Data ($N = 10,122$) in a first step. The probability of both partial unit nonresponse and on item nonresponse is regressed on age, sex, education, nationality, number of adults, number of children, big region (NUTS-2 level), urbanicity, home ownership, and a number of deprivation variables (holidays, car, computer, and noisy flat). Missing values in these explanatory variables were imputed using chained equations.

The McFadden pseudo chi^2 of the MAR unit nonresponse model amounts to .079, that of the item-nonresponse model to .025. The NMAR mechanism in the SILC data is estimated adding income deciles from registry data to the explanatory variables. The McFadden pseudo chi^2 of the unit-missing model amounts to .090, of the item missing model to .030. Our analysis suggests an NMAR nonresponse mechanism for partial unit nonresponse and a MAR mechanism for item nonresponse: after controlling for observed independent variables, while the unit nonresponse probability depends on the joint categories of employment income, the item nonresponse probability does not. We apply both the estimated nonresponse models for a MAR and an NMAR mechanism from the SILC-data to the SHP data, using the same independent variables. Since the results turn out to be comparable, we decided to focus on the estimated NMAR mechanism. We add the longitudinal information of unit nonresponse and item-nonresponse in the previous wave (Westermeier & Grabka, 2016), which is only available in the SHP data. Specifically, we first predict nonresponse probabilities in the SHP from the maximum of the SILC unit nonresponse probabilities and the previous wave unit missing probabilities, and second the maximum of the SILC item nonresponse probabilities and the previous wave item missing probabilities. Finally, to choose the observation to be set to missing, we assign a random number to each observation and set the case to missing if that random number is below the predicted nonresponse probability. This process is done separately for item- and unit nonresponse within the same data set. We first set the values to missings according to the unit nonresponse mechanism, then by the item nonresponse mechanism.

4.3 Implementation of the Imputation

We compare four approaches to deal with missing income data: iterative regression (1), LS (2), the mean and within approach (3), and complete-case analysis (4). For all imputation methods, we use real income to account for inflation. Both the iterative regression imputation and the mean and within approach are based on regression. For the iterative regression approach, we impute the independent variables including (logarithm of) income and lagged (logarithm of) income using chained equations. The exponent of the regression prediction is used as imputed value. For the mean and within approach, the same covariates (but excluding lagged income) are used. We first impute the independent variables also using chained equations before we regress (the logarithm of) income on the independent variables. The regression prediction is used to compute the "within" component, with exponentiation in the last step to calculate the within component.

For the LS approach, we implemented the standard LS method. Specifically, we use the LS imputation without stratification variable, and a new donor for each single imputation (i.e., not the same donor for all imputations of a respondent). All analyses were done using Stata 16 SE and ice algorithm for iterative regression imputation.

5 Results

To evaluate different imputation methods, we discuss performance under the empirically estimated NMAR nonresponse mechanism. For ease of readability, we depict graphs showing bias in the main text and include tables with the estimatesin the appendix.

[6] As the SHP, the SILC survey has grid, household and individual questionnaires and relies mostly on the telephone mode. The SILC survey collects employment income from calendar years, so information from the survey and the registry are comparable. Due to the highly comparable design of the two data sources, most SILC-respondents are no longer asked about their employment income as this information is used from the register.

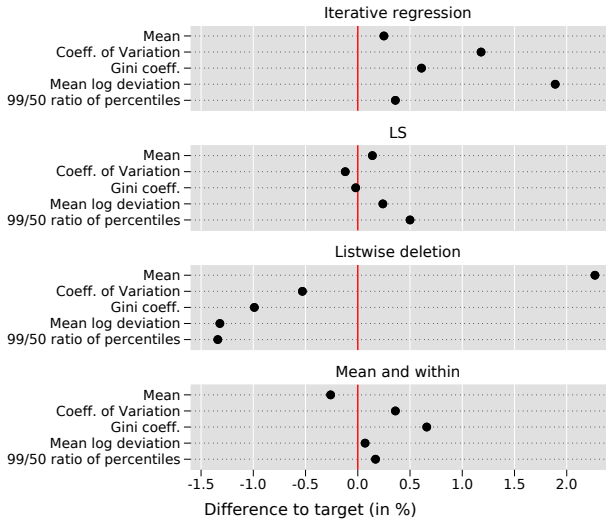

Figure 1

Relative difference to the target values. Target values are: Mean 27,123 CHF; Coefficient of variation 0.83; Gini 0.41; Mean log deviation 0.45; 99/50 ratio of percentiles 3.79; Data: SHP 2002–2021, $N = 102,891$ for imputed data, $N = 87,445$ for listwise deleted.

5.1 Cross sectional criteria

Descriptive statistics

We start with the estimation accuracy of the imputed data assessed with the mean and inequality measures; see Table A8 in the appendix. Figure 1 compares distributive statistics from observed and imputed data, showing the difference in relative terms. The closer to zero, the more accurate the imputation.

Starting with mean income, we see that all imputation procedures improve estimation compared to complete case analysis, which overestimates mean income by 2%. The three imputation methods are not significantly different from the target mean value.

For inequality measures, all imputation methods tested tend to decrease nonresponse bias. The LS estimates are unbiased for each of the indicators, as they do not differ significantly from the target value. Imputation by iterative regression slightly overestimates inequality for the Gini coefficient and the MLD. Imputation with the mean and within method slightly overestimates inequality for the Gini index. In contrast, complete case analysis tends to underestimate inequality for the Gini index and the MLD. Overall, using the average bias over the four inequality statistics, the LS methods performs clearly best (0.2% difference), followed by the mean and within approach (0.3%), iterative regression imputation and complete case analysis (both 1%).

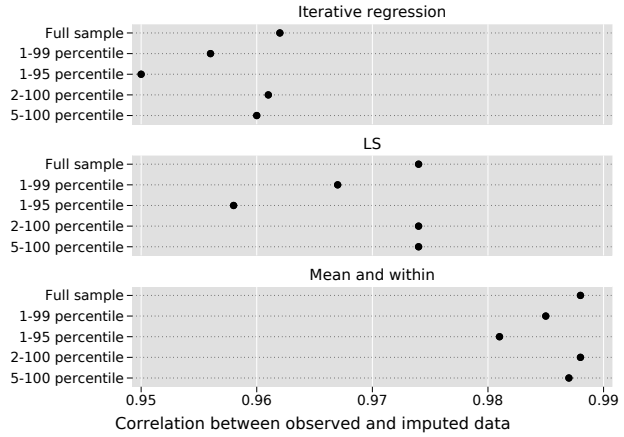

Figure 2

Correlations between observed and imputed values with and without extreme percentiles. Target value is 1. Data: SHP 2002–2021, $N = 102,891$.

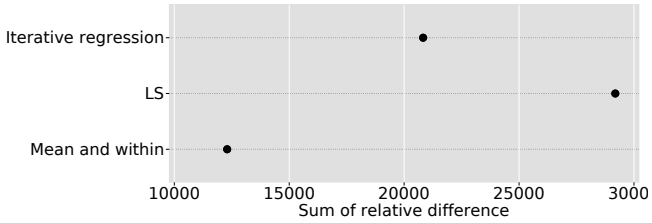

Figure 3

Sum of relative differences of imputed from observed values. Target value is 0. Data: SHP 2002–2021, $N = 102,891$.

Predictive accuracy 1

The correlations between observed and imputed values are calculated for the overall sample and samples excluding different percentiles at the lower or upper tails. Results are shown in Figure 2, with higher correlation reflecting lower bias (as the target value is 1); also see Table A1. Prediction in terms of correlations between observed and imputed values performs best with the mean and within approach, which outperforms both iterative regression and LS in all scenarios.

Predictive accuracy 2

The sum of the absolute relative differences to the target values[7] are shown in Figure 3, where lower values reflect better accuracy; also see Table A2. The difference between observed and imputed values is largest in the LS method compared to iterative regression and mean and within deviation, which performs best.

[7] $\sum \frac{\text{abs(reported-imputed)}}{\text{reported}}$

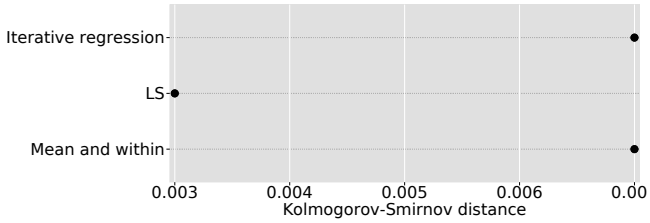

Figure 4

Maximum Kolmogorov-Smirnov difference between imputed and reported data. Target value is 0. Data: SHP 2002–2021, $N = 102,891$.

Kolmogorov-Smirnov distance

The Kolmogorov-Smirnov test refers to the maximum distance between the distributions of the reported and the imputed values is shown in Figure 4, with lower values indicating better accuracy; see Table A3. For iterative regression and the mean and within approach, differences between observed and imputed values are significant (5%). The LS method has the closest distribution of employment income to reported values (no significant difference).

5.2 Longitudinal criteria

Cross-wave correlations

The first longitudinal criterion is the correlation between income and lagged income, assessing whether the imputation captures stability over time. The target value is the cross-wave correlation in the reported data (amounting to 0.91). As for predictive accuracy (criteria 2), we both include and drop extreme percentiles. Figure 5 shows the (relative) deviation from the target value; also see Table A4. Note that this statistic can only be estimated for consecutive observations.

Interestingly, the tested imputation methods and complete case analysis tend to underestimate the true cross-wave correlation,[8] although the mean and within imputation and complete-case analysis shows only small bias, with average differences to the target amounting to 0.2%. Stability is more strongly underestimated in the LS approach (2%) and using iterative regression (5%).

Income mobility

The second longitudinal criterion compares income mobility, using income deciles between waves. We estimate both income mobility over one year and over five years. Figure 6 shows the relative deviation of the Spearman's rank correlation resulting from the different imputation methods to the target value (target $r = 0.91$ over 1 wave and $r = 0.76$ over 5 waves); see Table A5. Negative values indicate overestimation of income mobility, positive values underestimation of income mobility. Note again that this measure can only be computed for individuals with observations in both waves.

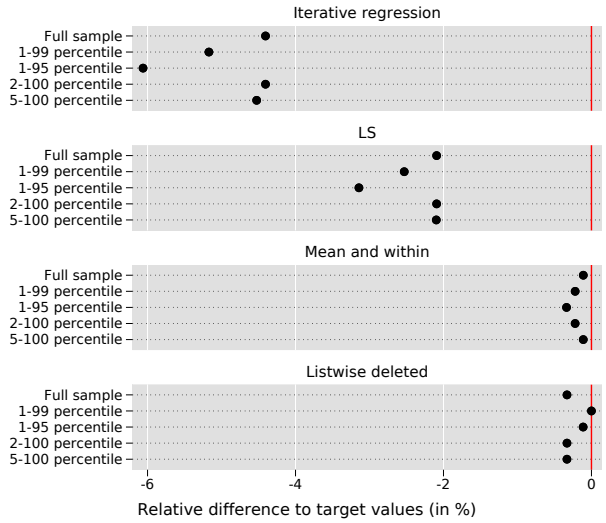

Figure 5

Correlation between income and lagged income for different percentile groups. Data: SHP 2002–2021, $N = 102,891$ *for imputed data,* $N = 87,445$ *for listwise deleted.*

The bias is strongest for iterative regression, where income mobility is relatively strongly overestimated (by 3 and 5%). The LS method also overestimates mobility over one wave (by 3%) but underestimates mobility over 5 waves (by 2%). The mean and within approach is close to the target (bias of 0.3% for each measure). The complete case analysis is unbiased for mobility over one year and slightly underestimates mobility over five waves (by 1%).

Between- and within-individual variation of income

Figure 7 reports the standard deviation of income between individuals and within individuals over time, again as relative deviation from the target value; see Table A6. LS underestimates within variation (by 3%) and is close to the target for between variation (0.4%). The mean and within imputation also slightly underestimates within variation and is also close for between variation with bias amounting to 1% and 0.3%, respectively. The iterative regression gives good results for between variation (bias of 0.4%), but leads to a considerably overestimation of within variation (12%). Complete case analysis underestimates within variance by 1%.

[8] The last value carried forward method overestimates stability, as could be expected (see online appendix).

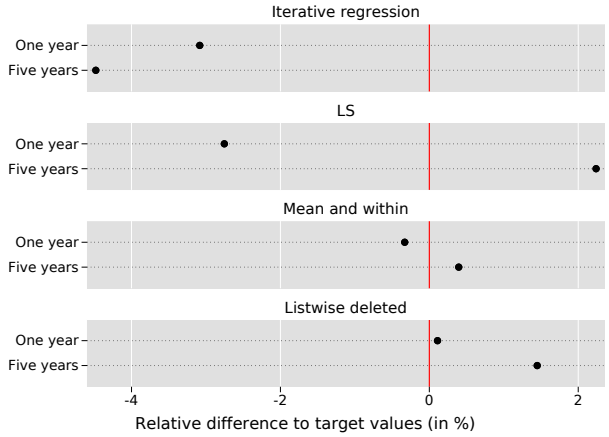

Figure 6

Spearman's rank correlation between lagged and current income deciles. Target value is 0. Data: SHP 2002–2021, N = 74,789 (one year lag); 40,061 (five years lag); 54,836/29,079 (listwise deleted).

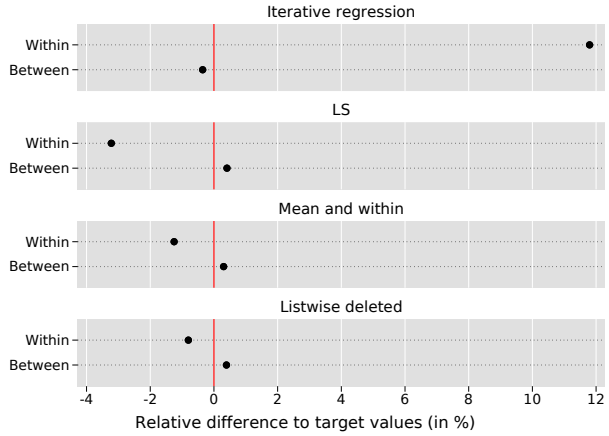

Figure 7

Between- and within standard deviation of income. Target values are: Within standard deviation 19,675 CHF; Between standard deviation 48,353. Data: SHP 2002–2021, N = 102,891 for imputed data, N = 87,445 for listwise deleted.

Application in exemplary regression models

As a final evaluation criterion, we test the imputed variables in a typical panel data application, namely pooled OLS regression ("cross-sections") and fixed effects ("within") models. In this example, we estimate the impact of employment income on life satisfaction among the working population. As dependent variable, we use satisfaction with life (measured on a 0–10 scale) and control for a large number of standard independent variables based on literature on this topic.[9] Employment income is included as a linear variable in units of 100,000 CHF and top coded at 99%.

We compare the regression coefficient for employment income on life satisfaction for the different methods of treating missing data. We do this both for the complete data where 15% of income values were imputed ($n = 102,866$), and the subsample with imputed values for income only ($n = 15,443$).[10] The later highlights the difference between the imputation methods, which become more important the higher the share of missing data is.

In the OLS regression, we estimate that an income increase of 100,000 per year is associated with an increased life satisfaction by 0.20 points (target value using reported data). The results are very close in all different models.

In the fixed effects regression, the estimated effect of income is smaller and amounts to 0.07, but confirms significant positive impact of income on life satisfaction. All estimated coefficients do not differ significantly from the target value and fall into the confidence interval for reported data (from 0.03 to 0.10). Therefore, imputation does neither improve nor bias imputation models compared with listwise deletion, but has the advantage of more power.

Focusing on the size of the bias, Figure 8 (Table A7) shows the relative difference of regression coefficients between the models with observed data (no missing values) and the models containing missing information, reflecting the bias resulting from nonresponse and imputation. For OLS regression, the complete case approach works best (bias of 1%), followed by mean and within (2%), LS (2%) and iterative regression (2%). For FE regression, the complete-case analysis works best (bias of 7%), while the bias of the three imputation methods is comparable (around 20%).

If we only consider only imputed values (Figure 9 and Table A7), the differences between the regression are amplified. For OLS regression, the mean and within imputation results in an overestimated income effect by 13%, although the difference is not significant. The iterative regression and LS understate the impact of income on life satisfaction by

[9] Survey year (2002–2021), sample (original, refreshment), region (NUTS II), relation to household head, survey status (respondent, partial unit non-respondent), age (6 groups), civil status (married vs. not), nationality (Swiss vs. not), partner, working status (full-time, part-time, mini job, retired, unemployed), number of adults in household, number of children in household, years of education, gender, parents have higher education, health, number of rooms, can make holiday, can invite friends, can go to restaurants, have third pillar, can go to dentist, live in noisy environment, household has arears of payments, household can save, household spends what it earns, household eats savings.

[10] The sample size is slightly lower than in the other evaluation criteria, due to missing values in life satisfaction (dependent variable).

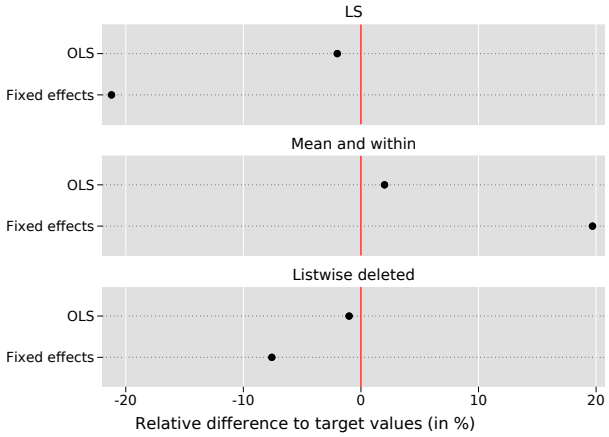

Figure 8

Errors of regression coefficients of income in longitudinal models (15% missing/imputed data). Target values are 0.07 for fixed effects regression and 0.20 for ordinary least square regression. Data: SHP 2002–2021, N = 102,866.

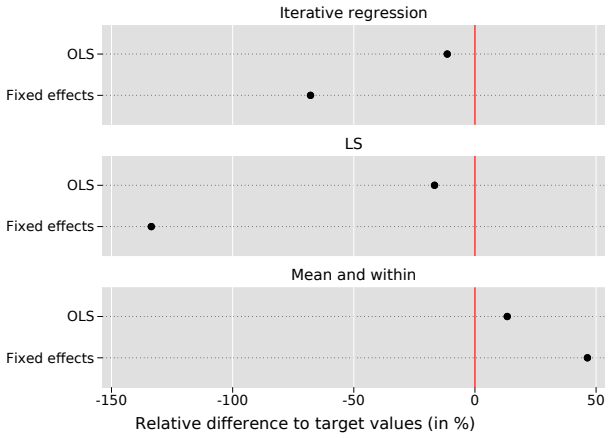

Figure 9

Errors of regression coefficients of income in longitudinal models (imputed data only). Target values are 0.14 for fixed effects regression and 0.21 for ordinary least square regression. Data: SHP 2002–2021, N = 15,443.

12 and 17%, respectively. For FE regression, in particular the LS imputation has strongly biased estimates, which differ significantly from the target. Instead of the positive income effect on life satisfaction (coefficient of 0.14 with reported data), LS imputation and iterative regression would wrongly suggest no effect of income on life satisfaction. For mean and within imputation the bias is smaller, does not differ significantly from the target value and shows a positive effect of income on life satisfaction.

5.3 Summary of results

To give an overview of the evaluation criteria, we computed the average bias for different sub-indicators (e.g., different inequality measures), and ranked the methods (see Table 2). It should be noted that this crude measure does not take the relevance of the bias and the closeness of the performance of the different imputation approaches into account. Its sole purpose is to provide a summary view.

For cross-sectional criteria, the LS method performed best, followed by the mean and within approach and finally iterative regression. The LS imputation is closest to the target values for criteria 1 (descriptive statistics) and 4 (Kolmogorov-Smirnov distance), but more biased for criteria 2 (correlation between observed and imputed values) and criteria 3 (mean absolute deviation from the income distribution), where mean and within approach performed best. Considering that mean and within approach relies on additional information (from covariates) compared to the univariate approach of the LS method, a better match between imputed and observed values for the accuracy measures is not surprising. The bad performance of "simple" iterative regression imputation in most criteria—which also relies on information of covariates—is more surprising. The very good performance of the LS approach in descriptive statistics (criteria 1) and 4 (Kolmogorov-Smirnov distance), which are both related to the distribution of income, confirms the conclusion of previous studies, that the LS is very well-suited approach for researchers who want to study income distribution. Complete case analysis, in general, perform worse than imputed values for cross-sectional analysis.

For longitudinal analyses, the LS imputation does not perform very well. For each longitudinal criteria, one of the other approaches worked better. The mean and within approach performed best for income mobility and cross-wave correlation, and better than LS and iterative regression for within variation. The iterative regression also falls behind the other imputation methods when it comes to longitudinal criteria. The complete case analysis performs relatively well, in particular for cross-wave correlation and within variation, where it shows the best results.

For the multivariate regression models, the conclusions are rather different. Here, none of the imputation methods tested improved the estimates from complete case analysis. Depending on the share of missing data, the estimates in fixed effects regression differ significantly from the target and could lead to wrong conclusions. This is the case for LS and iterative regression analysing observations with missing income data.

In addition to the methods presented here, we also tested last value carried forward (carry-over) imputation. This method performed quite well for both cross-sectional and longitudinal analysis, but did not outperform LS in the cross-sectional criteria and mean and within and complete case in

Table 2

Ranking of imputation methods

	Iterative regression	LS	Mean and within	Complete case
Cross sectional	3	1	2	4
1 Descriptive statistics				
a. Mean	2	1	3	4
b. Inequality measures	3	1	2	4
2 Correlation between observed and imputed values	3	2	1	-
3 Absolute deviation from observed values	2	3	1	
4 Distribution: Kolmogorov-Smirnov	2	1	3	-
8a Multivariate: OLS regression	4	3	2	1
Sum (Cross sectional)[a]	16	11	12	-
Sum wihtout multivariate	12	8	10	-
Longitudinal	4	3	1	2
5 Comparison of cross-wave correlations of the true and the imputed values	4	3	1	2
6 Income mobility (in deciles) in the imputed data compared with the one with observed values	4	3	1	2
7 Variability over time				
a. within individuals	4	3	2	1
8b Multivariate: FE regression	3	4	2	1
Sum (longitudinal)	15	13	6	6
Sum without multivariate	12	9	4	5

[a] We did not include between variation, as we considered the coefficient of variation as part of inequality measures.

the longitudinal criteria. However, for the multivariate regression carry over performed best in both the OLS and FE model.

Moreover, we also tested variants of the mean and within imputation: the estimation of the within-component by FE regression, and the estimation of the row effect (as used in LS) instead of the mean effect. Both variants yielded overall consistent conclusions, with sometimes the variants performing slightly better or slightly worse than the selected mean and within approach.[11]

6 Conclusion

This contribution tests different imputation methods for income in longitudinal data. As panel data is designed for longitudinal data analysis, it is important to consider longitudinal aspects when implementing an imputation strategy. The imputations are evaluated on various cross-sectional and longitudinal performance criteria for employment income using 20 annual waves from the Swiss Household Panel. We extend these criteria by testing the imputation methods in a typical multivariate regression model, comparing the effect of employment income on life satisfaction.

The first method assessed is the LS method, which is widely used in household panel data and relatively simple to apply. In line with the previous literature, we find that the LS method works well for cross-sectional application, in particular population averages and income distribution measures. The LS performs less for longitudinal criteria, where cross-wave correlation tends to be underestimated, mobility over one year overestimated and mobility over five years underestimated. We suspect that ignoring the systematic variation within individuals in the imputation model explains the bias of the LS method in longitudinal applications.

The second method assessed was iterative regression including lagged income. Regression-based approach allow to consider additional variables (besides income) in the imputation, but also add complexity compared to the LS approach. Despite of this, the iterative regression underperforms LS in both cross-sectional and longitudinal criteria. It needs to be noted that our iterative regression model was relatively simple. The approach could be enhanced, for instance, by additional time lags or implementing an algorithm accounting for the multilevel structure, but it remains to be tested whether

[11] The FE variant performs slightly better for criteria 1a, 4, 7, 8 and slightly worse for criteria 1b, 3, 6 and no difference for criteria 2, 5. The row&within variant performs slightly better for criteria 3 and slightly worse for criteria 1, and no difference for criteria 2, 5, 6,7, 8.

the complexities and additional assumptions needed actually improve the performance.

As a third approach, we tested an alternative for longitudinal data, which aimed to combine the advantages of the LS method and regression-based approaches. Based on individual-specific mean values (mean component), individual-specific deviations are included using regression models (within component). The mean and within imputation works nearly as well as the LS approach for cross-sectional analysis. For longitudinal analysis, the bias is considerably smaller than for LS and iterative regression, though. The mean and within approach seems particularly well suited to estimate income mobility. There seems to be a small tendency of regression to the mean for within analysis, which could be addressed in further studies with an additional random component.

In addition to evaluation criteria used in previous studies, we tested the imputations in a multivariate regression analysis as an application. Interestingly, the imputation did not improve results from multivariate regression compared to complete cases analysis other than increasing power in terms of sample size. While imputing missing income seems important for descriptive statistics, including control variables (related to nonresponse) can correct for nonresponse bias in multivariate models.

In a nutshell, we find that LS can be recommended when data are used cross-sectionally (such as income inequality, population averages or cross-sectional regression models), as it is simple and transparent, robust and well-suited for comparative analyses. For longitudinal data analysis, the mean and within approach, the carryover method or even complete case analysis propose smaller bias. In particular, for researchers who study income mobility, the mean and within approach could be an alternative to LS methods that merits to be further investigated. For multivariate regression, the imputation method is important only if a large share of data was imputed.

The different performances of imputation methods largely depend on the criteria applied. This rises the question of the usefulness and problems of providing imputed ("all-purpose") data for data users. While imputed values should be based on an adequately specified imputation model, it is problematic to shift the responsibility of imputing data to the data user from a practical perspective: first, only few users are experts for imputation and, second, it is in the interest of transparent and replicable research that researchers are able to work with the same imputed data (Axenfeld et al., 2022). Although providing an imputed dataset to the scientific community is highly valuable, data providers should make it transparent for which purposes imputed values are recommended and for which purposes caution and probably alternative imputations are needed. Ideally, researchers should reflect on and implement an imputation strategy targeted to their analysis.

As the quality of different imputation approach depends on the response mechanism, type of data and type of analysis, caution is required from inferring findings from specific studies on other type of data. Our study differs from most simulation studies though, as it does not only rely on real data that researchers use to estimate income distribution and income mobility, but also established the response mechanism empirically through linkage of survey and registry data from an external (comparable) data source. This allowed to detect and apply a NMAR response mechanisms, where response probability depends on income. Ideally, one would have true (administrative) income measures for all observations from a single data source. This would avoid the assumption of the same response mechanism in different surveys of differently concerned samples.

This study is only one step to developing better imputation methods and to establish best practices for longitudinal analysis. The tests should be extended to additional criteria, other longitudinal data, other assumptions on nonresponse mechanism, and other application examples for longitudinal models. Furthermore, the imputation should also be tested for income variables other than income from employment. It is likely that variation in income from employment within individuals over time can be more easily predicted than variation in income from other sources, as the survey contains many job-related information. In addition, reflections on how the approach can be extended to multiple imputation are required.

Despite the clear need for further tests and development of the method, we think that our paper provides an important first step towards new ideas to improving the imputation methods for longitudinal data analysis. In any case, researchers should be critical about imputed values provided with the data when conducting longitudinal analysis.

7 Acknowledgements

This study has been realized using the data collected by the Swiss Household Panel (SHP), which is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is financed by the Swiss National Science Foundation.

References

Andreß, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied panel data analysis for economic and social surveys*. Springer Science & Business Media.

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F., & Bela, A. (2017). Nonparametric multiple imputation for questionnaires with individual skip patterns and constraints: The case of income imputation in the National Educational Panel Study. *Sociological Methods & Research*, *46*(4), 864–897.

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., van Buuren, S., & Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, *33*(2), 160–183.

Axenfeld, J. B., Bruch, C., & Wolf, C. (2022). General-purpose imputation of planned missing data in social surveys: Different strategies and their effect on correlations. *Statistic Surveys*, *16*, 182–209.

Champney, T. F., & Bell, R. (1982). Imputation of income: A procedural comparison. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 431–436.

De Luca, G., Celidoni, M., Trevisan, E., et al. (2015). Item non response and imputation strategies in SHARE wave 5. In F. Malter & A. Börsch Suppan (Eds.), *SHARE wave 5: Innovations and methodology* (pp. 85–100). Munich Center for the Economics of Ageing.

Durrant, G. B. (2009). Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. *International journal of social research methodology*, *12*(4), 293–304.

Frick, J. R., & Grabka, M. M. (2014). *Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution* [SOEP Survey Papers].

Frick, J. R., Jenkins, S. P., Lillard, D. R., Lipps, O., & Wooden, M. (2007). The cross-national equivalent file (CNEF) and its member country household panel studies. *Journal of Contextual Economics—Schmollers Jahrbuch*, (4), 627–654.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Little, R. (1992). Regression with missing X's: A review. *Journal of the American statistical association*, *87*(420), 1227–1237.

Little, R., & Su, H. (1989). Item non-response in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton, & M. Singh (Eds.), *Panel surveys* (pp. 400–425). John Wiley; Sons, New York.

Spiess, M., Kleinke, K., & Reinecke, J. (2021). Proper multiple imputation of clustered or panel data. In P. Lynn (Ed.), *Advances in longitudinal survey methodology* (pp. 424–446). Wiley. https://doi.org/10.1002/9781119376965.ch17

Tillmann, R., Voorpostel, M., Kuhn, U., Lebert, F., Ryser, V.-A., Lipps, O., Wernli, B., & Antal, E. (2016). The Swiss household panel study: Observing social change since 1999. *Longitudinal and Life Course Studies*, *7*(1), 64–78.

Watson, N., & Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics*, *27*(4), 693.

Westermeier, C., & Grabka, M. M. (2016). Longitudinal wealth data and multiple imputation: An evaluation study. *Survey Research Methods*, *10*(3), 237–252.

Appendix
Tables

Table A1

Correlations between observed and imputed data

	Full sample	1–99 percentile	1–95 percentile	2–100 percentile	5–100 percentile
Iterative regression	0.962	0.956	0.950	0.961	0.960
LS	0.974	0.967	0.958	0.974	0.974
Mean and within	0.988	0.985	0.981	0.988	0.987
Carryover	0.986	0.983	0.979	0.986	0.986
Mean and within with FE	0.989	0.986	0.983	0.988	0.988
N	102,891				

Table A2

Sum of relative differences of imputed from observed values

	Sum of differences
Iterative regression	20,827
LS	29,188
Mean and within	12,301
Carryover	13,715
Mean and within with FE	12,610
N	102,891

Table A3

Maximum difference between imputed and reported data (Kolmogorov-Smirnov distance)

	Estimate	p-value
Iterative regression	0.007	0.02
LS	0.003	0.73
Mean and within	0.007	0.01
Carryover	0.005	0.18
Mean and within with FE	0.006	0.08
N	102,891	

Table A4

Correlation between income and lagged income for different percentile groups

	Full sample	1-99 percentile	1-95 percentile	2-100 percentile	5-100 percentile
Reported (target)	0.908	0.909	0.891	0.908	0.906
Iterative regression	0.868	0.862	0.837	0.868	0.865
LS	0.889	0.886	0.863	0.889	0.887
Mean and within	0.907	0.907	0.888	0.906	0.905
Listwise deleted	0.905	0.909	0.890	0.905	0.903
Carryover	0.915	0.918	0.902	0.915	0.914
Mean and within with FE	0.909	0.910	0.893	0.909	0.907
N	102,891				

Table A5

Spearman's rank correlation between lagged and current income deciles

	Lag 1 year	Lag 5 years
Reported (target)	0.908	0.759
Iterative regression	0.880	0.725
LS	0.883	0.776
Mean and within	0.905	0.762
Listwise deleted	0.909	0.770
Carryover	0.917	0.767
Mean and within with FE	0.909	0.772
N	74,789	40,061
N (listwise deletion)	54,836	29,079

Table A6

Between- and within standard deviation

	Within	Between
Reported (target)	19 675	48 353
Iterative regression	21 996	48 182
LS	19 041	48 551
Mean and within	19 429	48 500
Listwise deleted	19 517	48 543
Mean and within with FE	19 093	48 529
N	102,891	
N (listwise deletion)	87,445	

Table A7

Errors of regression coefficients of income in longitudinal models

	15% missing income				100% missing income			
	OLS estimate	p	FE estimate	p	OLS estimate	p	FE estimate	p
Reported	0.199		0.066					
Iterative regression	0.195	0.144	0.053	0.195	0.186	0.308	0.045	0.091
LS	0.195	0.163	0.052	0.213	0.175	0.133	−0.047	0.027
OLS mean and within	0.203	0.088	0.079	0.097	0.238	0.122	0.205	0.209
Listwise deleted	0.197	0.505	0.061	0.624				
Carryover	0.197	0.300	0.056	0.225	0.187	0.208	0.096	0.477
OLS row and within	0.202	0.124	0.080	0.076	0.226	0.383	0.218	0.138
FE mean and within	0.202	0.224	0.070	0.579	0.234	0.183	0.192	0.370
N	102,866				15,443			

p refers to p-value of Chi-square test of difference to estimates with reported data.

Table A8

Descriptive statistics

		95% C.I.	
	Estimate	Lower	Upper
Mean			
Reported (target)	57,123	56,833	57,414
Iterative regression	57,269	56,974	57,563
LS	57,204	56,914	57,494
Mean and within	56,976	56,686	57,267
Listwise deleted	58,419	58,098	58,739
Carryover	57,013	56,723	57,303
Mean and within with FE	57,033	56,743	57,323
COV			
Reported (target)	0.831	0.806	0.857
Iterative regression	0.841	0.815	0.867
LS	0.830	0.805	0.856
Mean and within	0.834	0.809	0.860
Listwise deleted	0.827	0.798	0.856
Carryover	0.833	0.808	0.859
Mean and within with FE	0.832	0.806	0.858
Gini			
Reported (target)	0.413	0.411	0.415
Iterative regression	0.415	0.413	0.417
LS	0.413	0.410	0.415
Mean and within	0.415	0.413	0.418
Listwise deleted	0.409	0.406	0.411
Carryover	0.414	0.412	0.416
Mean and within with FE	0.414	0.440	0.449
MLD			
Reported (target)	0.450	0.445	0.454
Iterative regression	0.458	0.453	0.463
LS	0.451	0.446	0.455
Mean and within	0.450	0.445	0.455
Listwise deleted	0.444	0.438	0.449
Carryover	0.455	0.450	0.460
Mean and within with FE	0.444	0.440	0.449
99/50 ratio of percentiles			
Reported (target)	3.79	3.73	3.85
Iterative regression	3.80	3.74	3.87
LS	3.81	3.74	3.87
Mean and within	3.80	3.73	3.86
Listwise deleted	3.74	3.67	3.81
Carryover	3.81	3.74	3.87
Mean and within with FE	3.81	3.74	3.87
N		102,891	