*Centre for Statistical and Survey Methodology*

**The University of Wollongong**

**Working Paper**

08-12

# Imputation of Household Survey Data using Linear Mixed Models

Luise P. Lago and Robert G. Clark

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

# Imputation of Household Survey Data using Linear Mixed Models

Luise P. Lago [1]
*Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, Australia*
Robert G. Clark
*Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, Australia*

## Abstract

This paper investigates whether using a linear mixed model to impute missing values in household surveys leads to improvement over imputation using a linear model and other standard imputation methods. The mixed model imputes leads to clear although not large improvements in predictive accuracy and the estimation of means, standard deviations and deciles, particularly when non-response is informative.

*Keywords*: Household Surveys; Imputation; Income; Linear Mixed Model; Multilevel;

## 1. Introduction

Developing an efficient strategy for dealing with missing data is essential in the current climate of falling response rates (Yan, Curtin, & Jans, 2010) and increasing difficulty to make contact with households (Atrostic, Bates, Burt, & Silberstein, 2001). Missing data is undesirable as it can lead to bias and increased variance of point estimators (Haziza, 2009), as well as difficulty in applying standard analysis techniques, which often rely on complete data. Imputation is a typical post-survey strategy for dealing with missing data. An imputation model is formed to predict the unknown value based on other

[1] *Address for correspondence:* Luise Lago, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia.
E-mail: lago@uow.edu.au.

known data (see for example Groves & Couper, 1998, p15) with the aim to reduce nonresponse bias while allowing the dataset to be analysed as if it was complete.

This paper will focus on imputation in household surveys where more than one person in the household is selected, which are of particular interest as they raise the possibility of making use of one or more respondents within a household to impute its nonrespondents. The focus will be on all-per household designs where information is missing about a person from an otherwise fully responding household. This will include exploring the impact of explicitly accounting for the household structure in the imputation model for person-level item nonresponse in this setting.

Imputation methods will be considered for an outcome variable of interest, making use of a set of auxiliary variables known for the population and a set of explanatory variables, available for both respondents and nonrespondents in the household. For example, if a person from a responding household answers all (or most) questions except for personal income, an imputation model can be built for the outcome variable income based on a series of auxiliary variables such as state and remoteness and explanatory variables such as age, sex and labour force status.

The aim is to investigate imputation in a 2-level linear mixed model for people within housholds and compare to a single level approach. Imputation methods will only be considered for a continuous outcome variable. Section 2 describes the notation and contains a brief review, Section 3 details the imputation models considered, Section 4 describes a simulation of imputation under informative and non-informative missingness, and Section 5 contains results. Section 6 will draw conclusions and discuss areas for further investigation.

## 2. Notation and Review

### 2.1 *Notation*

Assume a sample is selected from a finite population $U$ of people of size $N$. The sample has households as the primary sampling unit and each in-scope person in the household is selected. In practice there may be an initial stage of selection of areas, but this will be ignored to concentrate on people within households where intra-cluster correlation is much higher and of more intrinsic interest. Let $s$ denote the sample of households with at least one respondent. Each household $j = 1, ..., m$ in $s$ consists of persons $i = 1, ..., N_j$. Let $n = \sum_{j=1}^{m} N_j$ be the sample size of people. A set of auxiliary variables $\mathbf{z_{ij}}$ are assumed known on each person in the population, a set of $p$ explanatory variables $\mathbf{x_{ij}}$ are assumed completely observed on each person in the sample and the outcome variable $Y_{ij}$ is observed only for responding people. The notation $\mathbf{Y} = (\mathbf{Y}_o, \mathbf{Y}_u)$ is used to segregate the outcome variable in the sample into item-respondents $\mathbf{Y}_o$ (observed) and item non-respondents $\mathbf{Y}_u$ (unobserved). Also let $\mathbf{X} = (\mathbf{X}_o, \mathbf{X}_u)$ be the matrix of explanatory variables representing the full respondents and partial respondents respectively. Let $I_{ij}$ be a sample selection indicator such that $I_{ij} = 1$ if person $ij$ is selected in the sample $s$ and 0 otherwise, and $R_{ij}$ indicate response status for outcome variable $Y$ for person $ij$ such that $R_{ij} = 1$ when $Y_{ij}$ is observed and 0 otherwise. Let $Y^*_{ij}$ be the imputed value of $Y_{ij}$ (when $R_{ij} = 0$) and $Y^*_{ij} = Y_{ij}$ when $R_{ij} = 1$.

### 2.2 *Missing Data Approaches*

Imputation methods are developed based on either implicit or explicit assumptions about the response mechanism, the process causing missingness. The missing data inference framework of Rubin (1987) describes the response mechanism in distinct classes: Missing At Random (MAR), Missing Completely At Random (MCAR) and not missing at random (NMAR). When the missing data mechanism is MCAR, $P(R_{ij}|Y_{ij}, I_{ij}, \mathbf{z}_{ij}, \mathbf{x}_{ij}) = P(R_{ij}|I_{ij})$ that is the response status is independent of both the observed and unobserved data. Under MAR, $P(R_{ij}|Y_{ij}, I_{ij}, \mathbf{z}_{ij}, \mathbf{x}_{ij}) = P(R_{ij}|I_{ij}, \mathbf{z}_{ij}, \mathbf{x}_{ij})$ and the response status is random after conditioning on the observed data. When the missing data mechanism is NMAR, the nonresponse status is dependent on the outcome variable in a way that can't be conditioned away by known variables. Imputation methods often assume the response mechanism is either MCAR or MAR. The issue then is identifying variables $\mathbf{z}$ and $\mathbf{x}$ that make this assumption true. In a household survey setting both the nonresponse model and the imputation model could reasonably be expected to depend on the household structure, but this is rarely explicitly built in to the models used in practice. The simulation study which will be described in Section 4 considers household level factors in both the nonresponse and imputation models.

Imputation methods can be grouped into several different types. Firstly the imputation method may be determinstic or stochastic. Deterministic methods always produce the same impute given a set of characteristics and stochastic methods have a random component.

The imputation method may also be designed to produce more than one impute. One method of producing multiple imputes is by repeated imputation, that is repeatedly applying a stochastic imputation method. Rubin (1987, p118-119) requires a set of conditions to be met for the multiple imputes to be considered 'proper' and the resulting inference to be valid. The imputes must be drawn from the posterior distribution of the missing data conditional on the observed, $P(\mathbf{Y}_o|\mathbf{Y}_u)$ which requires a model for both the data and the missing data mechanism.

An alternative approach to imputation is weighting. Weighting accounts for nonresponse by dividing the sample into adjustment cells, and adjusting the weight given to respondents by the inverse of the response rate in that cell (Little & Rubin, 1987, p55). Dealing with nonresponse via weighting can be much less resource intensive than imputation however it is inefficient because data collected from partial respondents are not used. When there is a high level of partial nonresponse, weighting can result in large data loss. Weighting is more typically used for dealing with unit nonresponse than item nonresponse.

### 2.3 *Imputation using a Single Level Linear Model*

Linear regression models are regularly used for imputing missing continuous items (see for example Little & Rubin, 1987, p44) under the assumption that the response mechanism is MAR. Even in household surveys where values of $Y$ for people in the same household are most likely correlated, the assumption of independent errors is common. A single level population model for continuous $Y$ is $Y_{ij} = \mathbf{x}_{ij}\beta + \epsilon_{ij}$ where $e_{ij}$ are i.i.d. $\sim N\left(0, \sigma_\epsilon^2\right)$ random variables. The Best Linear Unbiased Predictor (BLUP) under this model using the observed data, $\mathbf{Y}_o$ is:

$$\hat{Y}_{\mathrm{LM},ij} = \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} \tag{1}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X_o}^T\mathbf{X_o})^{-1}(\mathbf{X_o}^T\mathbf{Y_o})$ is the ordinary least squares estimate of $\boldsymbol{\beta}$

(Little & Rubin, 1987, p23), $Y^*_{ij} = \hat{Y}_{ij}$ when $R_{ij} = 0$ and $Y^*_{ij} = Y_{ij}$ otherwise.

### 2.4 *Extending the Linear Model*

Pfeffermann (1988) developed *augmented regression predictors* by incorporating an adjustment to a single level regression prediction to incorporate clustering. This work was extended to consider nonresponse in a longitudinal setting by Pfeffermann and Nathan (2001) using a combination of time series methods and mixed linear models. Each time point had an individual two-level linear mixed model which were 'connected' by specifying a model for the household and individual level residuals over time. An empirical study looked at imputing number of hours worked during the week preceding the interview. To overcome problems with convergence in model estimation and negative variance estimates under Iterative Generalised Least Squares (IGLS) model parameters were estimated using state space methods. The improvements of the proposed models found in a simulation study were not replicated in this empirical study. The reasons suggested were fit of the model no longer being 'perfect', small household sizes (most with just one person) and smaller sample size for parameter estimation.

Elbers, Lanjouw, and Lanjouw (2003) formulated a linear mixed model with random effects for geographic clusters of households to impute household expenditure for census data. A simulation study showed that the imputation performed best in large clusters of households but not so well for small cluster sizes. Data was at the household level, so modelling of people within households was not considered.

### 2.5 *Evaluating Imputation Methods*

While it is routine to consider whether an imputation strategy preserves univariate and multivariate population distributions (David, Little, Samuhel, & Triest, 1986 and Marker, Judkins, & Wingless, 2002), in a household survey setting there are additional considerations. An important evaluation criteria specific to household surveys is the intracluster correlation, or ICC. Preserving relationships at the household level may be of particular importance in a household survey, for example a survey collecting household income may aim to improve understanding of the varying income levels within a household. In this case realistic within household income patterns are crucial. It is undesirable for the imputation strategy to artificially weaken or strengthen the clustering of variables within households.

An important part of the imputation process is evaluation of the imputation strategy. Ideally the analysis model is pre-determined and the imputation method then can be evaluated simply by its ability to reproduce any complete data analysis. In Chambers (2001) this is termed ‘preservation of analysis’. A rigorous set of criteria were also developed in Chambers (2001) as part of the EUREDIT project to evaluate new techniques for editing and imputation. Five performance requirements for an imputation method are described: predictive accuracy, ranking accuracy, distributional accuracy, estimation accuracy and imputation plausibility. The first of these two criteria are noted to be of less relevance when estimates are of population aggregates, however for public release datasets and when the imputed data will be used in prediction models these criteria are of key importance. Three of Chambers' criteria were used to evaluate the imputation models in this paper:

(a) Predictive accuracy analyses the performance of the imputation model in reproducing the true values.

(b) Distributional accuracy evaluates the reproduction of the marginal distribution and moments of the distribution of the imputed data compared to the distribution of the true values.

(c) Estimation accuracy considers the performance of the imputation methods in reproducing low-order moments of the of the distribution of the true values which should then lead to unbiased estimates of parameters relating to the distribution of the true values.

Pfeffermann and Nathan (2001) used Relative Root Mean Square Error (RRMSE) and Relative Bias to compare the predictive accuracy of various imputation methods over $R$ replicates. The RRMSE after imputation can be calculated as follows:

$$\begin{aligned}\text{RRMSE}_{av} &= \frac{1}{R}\sum_{r=1}^{R}\text{RRMSE}_r \\ &= \frac{1}{R}\sum_{r=1}^{R}\left\{\sqrt{\frac{\sum_{ij\epsilon S_r}\left(y^*_{ij,r}-y_{ij}\right)^2}{\sum_{ij\epsilon S_r}(1-R_{ij,r})}}\Bigg/\frac{\sum_{ij\epsilon S_r}(1-R_{ij,r})\,y_{ij}}{\sum_{ij\epsilon S_r}(1-R_{ij,r})}\right\} \qquad (2)\end{aligned}$$

Relative bias is calculated for each replicate and averaged over the $R$ replicates:

$$\begin{aligned}\text{RBias}_{av} &= \frac{1}{R}\sum_{r=1}^{R}\text{RBias}_r \\ &= \frac{1}{R}\sum_{r=1}^{R}\frac{\sum_{ij\epsilon S_r}\left(y^*_{ij,r}-y_{ij}\right)}{\sum_{ij\epsilon S_r}y_{ij}(1-R_{ij})} \qquad (3)\end{aligned}$$

In Chambers (2001) the preservation of the distribution of a scalar variable is measured by categorising the distribution of true and imputed values, then assessing the proportion of imputes which have changed category.

## 3. Linear Mixed Model

When a variable of interest is considered likely to be correlated within households, linear mixed models can be used. A mixed model (Goldstein, 2003, West, Welch, & Galecki, 2007) considers the regression coefficients to be random variables with different realisations for each household. The two-level linear mixed model is

$$Y_{ij} = \beta_0 + \beta_1\mathbf{x}_{ij} + (u_{oj} + u_{1j}\mathbf{x}_{ij}) + e_{ij}.$$

So the regression coefficients can be expressed as $\beta_{0j} = \beta_0 + u_{0j}$ and $\beta_{1j} = \beta_1 + u_{1j}$, where $u_{0j}$ and $u_{1j}$ are random variables with $E(u_{0j}) = E(u_{1j}) = 0$, $var(u_{0j}) = \sigma^2_{u0}$, $var(u_{1j}) = \sigma^2_{u1}$ and $cov(u_{0j}, u_{1j}) = \sigma_{u01}$.

Households only contain a small number of people (often just one), so a special case, the random intercept model, is typically used. This restricts the random component to the intercept term only:

$$Y_{ij} = \beta_0 + \beta_1 \mathbf{x}_{ij} + u_{oj} + e_{ij}.$$

We will henceforth write $u_{0j}$ as $u_j$ for simplicity. We write $\boldsymbol{\beta}$ for the p-vector of regression coefficients for the fixed part of the model and $u_j$ and $e_{ij}$ are referred to as the household and person level residuals respectively.

Correlation of a continuous variable within households is measured by the Intra-Class Correlation (ICC). The ICC is defined as the proportion of total variation due to clustering within households, $ICC = \sigma^2_u/(\sigma^2_u + \sigma^2_e)$ (West et al., 2007, p98). This parameter is sometimes referred to as the "adjusted" ICC, because fixed effects $\mathbf{x}$ are included in the model, so that the ICC refers to the residual correlation after removing the effect of these variables. The unadjusted ICC is defined similarly but is based on a model where the fixed effects consist of an intercept only.

Clark and Steel (2002) found within household unadjusted ICCs in the range of 0.03 (full-time student) to 0.86 (English as a Second Language) with correlations typically between 0.1 and 0.3. If there is a strong correlation between individuals within a household there may be benefit in using the known values from respondents to impute non-respondents in the same household. Where there is little or no correlation there may be limited value in incorporating the household structure. In addition to potentially improving the accuracy of the imputed missing value, taking into account within household correlation will affect the resulting within-household correlation post-imputation compared to a person-level imputation model. If household structure is ignored in imputation then not only is potentially valuable information being disregarded, but the resulting imputes may distort within household patterns. Often household surveys are designed with a prime purpose of understanding household type attributes including aggregates of person level items. Therefore imputation should not only accurately reproduce univariate and multivariate relationships but also within household correlations.

Under a mixed linear model, a BLUP can be derived for the fixed and random effects and for the missing values $\mathbf{Y_m}$. The BLUP for predicting missing $Y_{ij}$ under this model can be shown to be the single level regression predictor plus a term incorporating the within household covariance:

$$\hat{Y}_{\mathrm{LMM},ij} = \hat{Y}_{\mathrm{LM},i,j} + C(\mathbf{y}_o, Y_{ij})\mathbf{V}_o^{-1}\{\mathbf{y}_o - \mathbf{x}_o(\mathbf{x}_o^T\mathbf{V}_o^{-1}\mathbf{x}_o)^{-1}(\mathbf{x}_o^T V_o^{-1}\mathbf{y}_o)\}$$

where $C(\mathbf{y}_o, Y_{ij})$ is a vector of covariances between the observed $\mathbf{y}_o$ and the missing value $Y_{ij}$ and $\mathbf{V}_o$ is a block diagonal matrix, with blocks $\mathbf{V}_{o,j} = \sigma^2((1-\rho)\mathbf{I}_{n_{o,j}} + \rho\mathbf{1}_{n_{o,j}}\mathbf{1}^T_{n_{o,j}})$ for $j = 1, ...m$, $n_{o,j}$ is the number of item respondents in household $j$, and $\mathbf{1}_{n_j}$ is a column vector of 1's of length $n_{o,j}$.

Assuming there is no covariance between people in different households this can be simplified to:

$$\hat{Y}_{\mathrm{LMM},ij} = \hat{Y}_{\mathrm{LM},i,j} + C(\mathbf{y}_{o,j}, Y_{ij})\mathbf{V}^{-1}_{o,j}\{\mathbf{y}_{o,j} - \mathbf{x}_{o,j}(\mathbf{x}^T_{o,j}\mathbf{V}^{-1}_{o,j}\mathbf{x}_{o,j})^{-1}(\mathbf{x}^T_{o,j}V^{-1}_{o,j}\mathbf{y}_{o,j})\} \tag{4}$$

Barroso, Bussab, and Knott (1998) derived a general form of (4), Henderson (1975) used a similar model for prediction in animal breeding, and Pfeffermann (1988) applied a variant of this model for simulated longitudinal household survey data. This paper specifically looks at cross-sectional household survey data.

In a household survey with nonresponse, the variance parameters $\sigma^2_u$ and $\sigma^2_e$ will be unknown and therefore need to be estimated. The predictions resulting from substituting estimates for these variance parameters is known as the empirical BLUP (Barroso et al., 1998). Several estimators are available, including Maximum Likelihood, Restricted Estimation by Maximum Likelihood (Patterson & Thompson, 1971), and Minimum Variance Quadratic Unbiased Estimation (Searle, Casella, & McCulloch, 1992). The first two of these methods are iterative techniques while the latter provides a non-iterative alternative with reduced processing time and not requiring normality (Wang, Xie, & Fisher, 2012).

## 4. Simulation study

### 4.1 *Imputation variable from HILDA*

A simulation study was carried out by applying a set of imputation models to a continuous variable from the Household Income and Labour Dynamics of Australia Survey, or HILDA (Watson, 2008). HILDA is an annual longitudinal survey which commenced in Australia in 2001. Hourly Wage Rate was the variable selected from Wave 4 of HILDA (2004) for the simulation study because income is a high priority for the survey and has high rates of partial non-response. Yan et al. (2010) found that income (and hence hourly wage rate which is derived from income) typically has high levels of nonresponse, of the order of 20-40% compared to other survey questions where item nonresponse was between 1 and 4%.

The sample was subsetted to people who were respondents to the data item hourly wage rate. This consisted of 4,820 persons in 3,318 households. Non-response could then be simulated and the various imputes compared to known values. As the imputation method is designed to make use of one or more respondents within the household, the sample was restricted to households with two respondents. This removes single respondent households and also allows comparison with another household-related method, imputing the nonrespondent within the household with the respondents value. This resulted in a sample of 2,392 persons from 1,199 households, representing approximately 50% of the responding sample.

### 4.2 *Simulating non-response*

Non-response was generated in $R = 250$ replicate samples taken from the fully observed component of the sample, to isolate the impact of the non-response mechanism and imputation method as distinct from population or sample variation. Half of households were designated to have nonresponse according to the different response models described below, and one of the two people within nonresponding households was selected to be a nonrespondent. This gives an overall response rate of 75%. Four alternative models were used to generate nonresponse. The first has data missing completely at random, and the others have informative nonresponse:

- **Households MCAR and persons MCAR:** Households were assigned randomly and independently to be fully responding or partially responding, with 50% probability of each. In the fully responding households, all variables were assumed to be collected for both household members. In the partially responding households, one person was randomly chosen to be the full respondent, and assumed to provide all variables. The other person was assumed to be partially responding, failing to provide income, but providing all other variables.

- **Households MCAR and persons NMAR** Again, households were assumed to have a 50% chance of being fully responding and a 50% chance of being partially responding. In partially responding households, one person was assumed to provide income data and the other wasn't. The probability of being the partial respondent within the household was inversely proportional to the person's hourly wage rate, i.e. lower wages were associated with higher response rate.

- **Households NMAR and persons MCAR** Households were assigned as partially or fully responding, with the probability of the household falling in the first category proportional to the household mean hourly wage rate. This was done so as to have approximately 50% of households fully responding. Within partially responding households, one randomly chosen person was chosen to be a full respondent, while the other was assumed to not provide income.

- **Households NMAR and persons NMAR** Households were assigned to be fully or partially responding in the same way as in the previous dotpoint. Within partially responding households, one person was assumed to partially respond and the other to fully respond, with the probability of partially responding being inversely proportional to the person's hourly wage rate.

### 4.3 *Imputation methods*

Six different imputation methods were compared in the simulation study for imputing missing $y_{ij}$ given a set of respondents $\mathbf{y_o}$, which includes a responding person $y_{i‘j}$ in the same household:

(a) **Respondent Mean:** $y^*_{ij}$ = mean of $Y$ over all fully responding people in the sample.

(b) **Household respondent:** the hourly wage rate for the other person in the household $y^*_{ij} = y_{i'j}$

(c) **Donor:** a random person selected from all respondents

(d) **Class donor:** a random person selected from respondents within the same agegroup by sex class

(e) **Single Level BLUP:** empirical BLUP for single level linear model, defined by equation (1).

(f) **Multilevel BLUP:** empirical BLUP for linear mixed model, defined by equation (4)

For the last two models a log transform of the outcome was performed prior to imputation, and the imputes back-transformed to be on the original scale with a bias correction (see for example David et al., 1986).

The BLUP imputation methods used the explanatory variables age by sex as these are available on the household form, and therefore are likely to be available for people in responding households regardless of whether the person themselves was a respondent. The ICC for hourly wage rate in the full sample was 0.194.

## 5. Results

The results comparing the BLUP under single-level and mixed linear imputation models for imputing *Hourly wage rate* are shown below. Respondent mean imputes, random donor imputes, within class donor imputes (Kalton & Kasprzyk, 1982) using age by sex to define classes and imputing using the household respondent were also calculated as a point of comparison.

Predictive Accuracy was assessed at an individual level by calculating the RRMSE of prediction as in equation (2), and the relative bias as in equation (3), averaged over the 250 replicates. Distributional accuracy was evaluated by considering the relative bias of deciles. Estimation accuracy was assessed for means, standard deviation and intra-household correlation (ICC) under the different nonresponse models for each imputation method, each averaged over 250 replicates.

Table 1 shows the predictive accuracy as measured by the RRMSE of each imputation method. For the first non-response model (uninformative at both household and person levels), the lowest RRMSE is achieved by the ML BLUP (55%), followed by the SL BLUP (56%) and the respondent mean methods (57%), with much higher values coming from the remaining methods. As expected, the RRMSE is higher for all non-donor methods when non-response is informative, particularly when both households and persons are NMAR. In all cases the ML BLUP achieves the lowest RRMSE, although it is never more than a few percent better than SL BLUP and respondent mean. The other three methods (household respondent, donor and class donor) have much higher RRMSEs. The RRMSE of the donor methods is not impacted when non-response is made informative.

Table 1 also shows that all methods have a negative bias when there is informative non-response. The household respondent method has small bias when only the household response is informative, but does poorly when person data is also NMAR. The ML BLUP imputation method generally has the least absolute bias when there is informative non-response. All methods have an appreciable bias under the fourth missingness model.

Compared to the SL BLUP, the ML BLUP has lower RRMSE in all four scenarios, and has less bias in all but the first scenario. The bias is 26%, 51%

and 23% less than that of the SL BLUP, under the second, third and fourth scenario, respectively.

Distributional accuracy was measured by the estimation of the highest two deciles of the distribution of hourly wage rate in Table 2. These top two deciles are where the nonresponse levels are highest, due to the NR model specifying that those persons and households where wage rates are lower are more likely to respond. The respondent mean and SL BLUP methods now perform the worst, underestimating the 8th and 9th deciles under each nonresponse model, including MCAR. While the donor respondent imputation methods perform poorly for predictive accuracy they have the least bias for estimating these top deciles. Using the other household member as a donor also leads to good reproduction of distributional accuracy. Although the ML BLUP does not result in bias as low as the donor and household respondent methods, it clearly out performs the respondent mean and SL BLUP. Compared to the SL BLUP the ML BLUP results in bias reductions of 92% and 12% for the 8th and 9th deciles respectively under MCAR, 50% and 3% reduction under persons NMAR, 66% and 9% reduction when households are NMAR, and 36% and 6% bias reduction in the 8th and 9th deciles respectively when both households and persons are NMAR.

Looking at estimation accuracy, Table 3 shows that the imputation methods are generally reasonably good for reproducing the mean, with all methods having a relative bias of less than 1.3% under MCAR and less than 9% for all other methods, including when both households and persons are NMAR. Under informative nonresponse models, the household respondent and ML BLUP imputes generally have the least bias for estimating the mean. There are more substantial biases in estimating the population standard deviation. The respondent mean and SL BLUP do poorly in reproducing standard deviation while the donor and household respondent imputes perform the best. ML BLUP imputes are worse for estimating standard deviation than the donor or household respondent methods, but a slight improvement on both the respondent mean and SL BLUP. Compared to SL BLUP, the ML BLUP imputes improve the relative bias for MCAR by 5%, when persons are NMAR by 3%, by 5% for households NMAR, and 3% improvement in relative bias is seen when both households and persons are NMAR.

Table 4 shows the relative bias in the estimated intraclass correlation due to imputation. The respondent mean, donor imputes and SL BLUP all underestimate the ICC, while the household respondent and ML BLUP impute overly similar values within a household. The household respondent method completely falls over, with very high levels of bias. This is due to all households with a nonrespondent having equal values of hourly wage rate and hence the ICC being severely over-estimated. The least bias for estimating the ICC results from ML BLUP under both MCAR and when persons are

NMAR. When households are NMAR none of the methods do a good job of reproducing the within household clustering.

Table 5 gives further information on the prediction accuracy. It shows how often the imputed value was in the same quartile as the true value for each person. ML BLUP is the best of the methods under all four response mechanisms. The improvement is not large, with a reduction in misclassification of about 5 percentage points over the single level BLUP.

## 6. Conclusions

The main question posed by this paper was whether imputations based on a two-level model for people within households do better than the more usual use of a single level model. The answer is yes, particularly when non-response is informative both of households and within households. The improvement is clear but not dramatic. The ML BLUP did slightly better in predictive accuracy, as measured by mean squared error, bias, and whether the impute was in the same quartile as the true value. The ML BLUP also reproduced the deciles of income more closely than the SL BLUP, particularly the 8th decile, where the relative bias under informative non-response at both stages was -14% and -9% for the ML and SL methods respectively. The two methods were very similar for the 9th decile. All imputation methods did similarly for estimating the overall mean, with low bias except in the most informative response mechanism. The single and multilevel methods both tended to under-estimate the overall standard deviation by about the same amount.

The multilevel BLUP might be expected to do well in reproducing the degree of within-household homogeneity, because it explicitly allows for dependencies within household. We found that the multilevel approach generally did better than the other imputation methods in the first three response scenarios. In the fourth scenario, the multilevel approach performed quite poorly in this regard. One possible reason is that both the single level and multilevel BLUP are deterministic imputes, and so understate the variability. The ML BLUP makes use of the dependency between a responding household member and their non-responding co-habitant. In the process, it tends to impute households which are too homogenous. In contrast, the SL BLUP tended to impute values which were too different from the other person in the household. Future research will use the multilevel model to create stochastic imputes which should give much more realistic levels of within-household homogeneity. These could be single imputes, or multiple imputes; the latter would incorporate imputation uncertainty into inference. The results of this paper suggest that this approach would be a useful improvement over single level imputation. Another avenue of future work will be to consider longitudinal household surveys by building in correlations across time as well as within households.

**Acknowledgements:** This research was funded by the Australian Research Council and the Australian Bureau of Statistics. Professors Ray Chambers and David Steel gave helpful comments and advice. The paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to any of the preceding organisations.

## References

Atrostic, B. K., Bates, N., Burt, G., & Silberstein, A. (2001). Nonresponse in U.S. Government household surveys: consistent measures, recent trends, and new insights. *Journal of Official Statistics*, *17*, 209-226.

Barroso, L. P., Bussab, W. O., & Knott, M. (1998). Best linear unbiased prediction in the mixed model with missing data. *Communs Statist. Theor. Meth.*, *27*(1), 121-129.

Chambers, R. L. (2001). *Evaluation Criteria for Statistical Editing and Imputation* (National Statistics Methodological Series No. 28). University of Southampton.

Clark, R. G., & Steel, D. G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, *70*, 289-314.

David, M., Little, R. J. A., Samuhel, M. E., & Triest, R. K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, *81*(393), 29–41. Available from `http://www.jstor.org/stable/2287965`

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

Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed.). Arnold.

Groves, R. M., & Couper, M. (1998). *Nonresponse in Household Interview Surveys.* John Wiley & Sons (New York; Chichester).

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications* (p. 215-246). Elsevier Science Publishers B. V.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2), 423-447.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-31.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley & Sons.

Marker, D. A., Judkins, D. R., & Wingless, M. (2002). Large-scale imputation for complex surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse* (p. 329-341). John Wiley & Sons Inc., New York.

Patterson, H. D., & Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, *58*(3), 545-554. Available from `http://www.jstor.org/stable/2334389` .

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *J. Am. Statist. Ass.*, *83*(403), 824–833.

Pfeffermann, D., & Nathan, G. (2001). Imputation for wave nonresponse - existing methods and a time series approach. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse* (p. 417-429). New York, USA, Wiley.

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

Searle, S., Casella, G., & McCulloch, C. (1992). *Variance Components.* John Wiley & Sons.

Wang, J., Xie, H., & Fisher, J. H. (2012). *Multilevel Models: Applications Using SAS.* Higher Education Press and Walter De Gruyter.

Watson, N. (2008). *Household Income and Labour Dynamics in Australia (HILDA) User Manual Release 6.* Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software.* Chapman & Hall/CRC.

Yan, T., Curtin, R., & Jans, M. (2010). Trends in income nonresponse over two decades. *Journal of Official Statistics*, *26*(1), 145-164.

Table 1: Predictive accuracy (relative root mean squared error % and relative bias %) for imputing *Hourly Wage Rate* (simulation standard errors shown in brackets)

| Imputation | $hh = MCAR$ | | $hh = MCAR$ | | $hh = NMAR$ | | $hh = NMAR$ | |
|---|---|---|---|---|---|---|---|---|
| method | $pers = MCAR$ | | $pers = NMAR$ | | $pers = MCAR$ | | $pers = NMAR$ | |
| | RRMSE | RBIAS | RRMSE | RBIAS | RRMSE | RBIAS | RRMSE | RBIAS |
| Resp Mean | 56.91 | 0.36 | 61.72 | -13.25 | 61.39 | -14.69 | 68.51 | -27.65 |
| | ( 0.34) | ( 0.17) | ( 0.29) | ( 0.14) | ( 0.27) | ( 0.14) | ( 0.16) | ( 0.10) |
| hh resp | 73.25 | 0.53 | 65.49 | -19.81 | 78.65 | 0.02 | 70.11 | -21.86 |
| | ( 0.29) | ( 0.19) | ( 0.25) | ( 0.13) | ( 0.15) | ( 0.20) | ( 0.12) | ( 0.12) |
| donor | 79.22 | -1.09 | 77.53 | -14.42 | 76.61 | -15.67 | 76.66 | -28.22 |
| | ( 0.33) | ( 0.21) | ( 0.25) | ( 0.19) | ( 0.26) | ( 0.17) | ( 0.15) | ( 0.13) |
| class donor | 77.57 | 0.50 | 75.37 | -11.98 | 75.25 | -12.92 | 75.26 | -25.45 |
| | ( 0.30) | ( 0.20) | ( 0.25) | ( 0.17) | ( 0.23) | ( 0.15) | ( 0.15) | ( 0.12) |
| SL BLUP | 55.51 | -0.22 | 60.50 | -12.47 | 60.16 | -13.83 | 67.17 | -25.91 |
| | ( 0.35) | ( 0.16) | ( 0.29) | ( 0.14) | ( 0.28) | ( 0.13) | ( 0.16) | ( 0.10) |
| ML BLUP | 54.84 | 5.07 | 58.61 | -9.16 | 58.56 | -6.77 | 64.72 | -21.13 |
| | ( 0.34) | ( 0.17) | ( 0.30) | ( 0.14) | ( 0.27) | ( 0.15) | ( 0.16) | ( 0.11) |

Table 2: Distributional accuracy for imputing *Hourly Wage Rate* - Relative Bias (%) of 8th and 9th Decile (simulation standard errors shown in brackets)

| Imputation | $hh = MCAR$ | | $hh = MCAR$ | | $hh = NMAR$ | | $hh = NMAR$ | |
|---|---|---|---|---|---|---|---|---|
| method | $pers = MCAR$ | | $pers = NMAR$ | | $pers = MCAR$ | | $pers = NMAR$ | |
| | $d8$ | $d9$ | $d8$ | $d9$ | $d8$ | $d9$ | $d8$ | $d9$ |
| Resp Mean | -7.81 | -8.06 | -10.97 | -11.31 | -12.07 | -11.60 | -14.63 | -16.16 |
| | (0.09) | (0.07) | (0.06) | (0.04) | (0.11) | (0.05) | (0.07) | (0.09) |
| hh resp | 1.77 | 0.45 | -3.03 | -5.87 | 1.64 | 0.15 | -4.33 | -7.93 |
| | (0.11) | (0.11) | (0.08) | (0.05) | (0.11) | (0.13) | (0.10) | (0.08) |
| donor | 1.14 | -0.29 | -1.13 | -4.98 | -1.74 | -5.31 | -7.13 | -9.98 |
| | (0.11) | (0.12) | (0.11) | (0.09) | (0.11) | (0.08) | (0.08) | (0.09) |
| class donor | 2.14 | 0.46 | -0.20 | -4.18 | -0.56 | -4.34 | -5.99 | -8.97 |
| | (0.11) | (0.13) | (0.05) | (0.09) | (0.08) | (0.09) | (0.11) | (0.09) |
| SL BLUP | -5.66 | -8.06 | -9.11 | -11.31 | -10.03 | -11.60 | -14.00 | -16.16 |
| | (0.08) | (0.07) | (0.09) | (0.04) | (0.06) | (0.05) | (0.05) | (0.09) |
| ML BLUP | -0.46 | -6.99 | -4.55 | -10.96 | -3.45 | -10.56 | -8.93 | -15.17 |
| | (0.05) | (0.09 ) | (0.07) | (0.06) | (0.07) | (0.08) | (0.08) | (0.08) |

Table 3: Estimation accuracy for imputing *Hourly Wage Rate* - Relative Bias (%) of Estimated Mean and Standard Deviation (simulation standard errors shown in brackets)

| Imputation | $hh = MCAR$ | | $hh = MCAR$ | | $hh = NMAR$ | | $hh = NMAR$ | |
|---|---|---|---|---|---|---|---|---|
| method | $pers = MCAR$ | | $pers = NMAR$ | | $pers = MCAR$ | | $pers = NMAR$ | |
| | $mean$ | $sd$ | $mean$ | $sd$ | $mean$ | $sd$ | $mean$ | $sd$ |
| Resp Mean | 0.1 | -13.3 | -3.7 | -19.8 | -4.1 | -20.0 | -8.7 | -32.7 |
| | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) |
| hh resp | 0.1 | 0.3 | -5.5 | -11.4 | 0.0 | -0.6 | -6.9 | -20.7 |
| | ( 0.0) | ( -0.3) | ( 0.0) | ( -0.2) | (-0.1) | ( -0.3) | ( 0.0) | ( -0.3) |
| donor | -0.3 | -0.8 | -4.0 | -7.9 | -4.4 | -8.4 | -8.9 | -22.9 |
| | ( 0.0) | ( -0.3) | (-0.1) | ( -0.3) | ( 0.0) | ( -0.3) | ( 0.0) | ( -0.3) |
| class donor | 0.1 | -0.4 | -3.3 | -7.6 | -3.6 | -7.6 | -8.0 | -21.8 |
| | ( 0.0) | ( -0.3) | ( 0.0) | ( -0.3) | ( 0.0) | ( -0.3) | ( 0.0) | ( -0.3) |
| SL BLUP | -0.1 | -12.6 | -3.5 | -19.1 | -3.9 | -19.4 | -8.2 | -32.1 |
| | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) |
| ML BLUP | 1.2 | -11.9 | -2.6 | -18.5 | -1.9 | -18.4 | -6.7 | -31.2 |
| | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) | ( 0.0) | ( -0.2) |

Table 4: Estimation accuracy for imputing *Hourly Wage Rate* - Relative Bias (%) of Estimated Intraclass Correlation (simulation standard errors shown in brackets)

| Imputation | $hh = MCAR$ | $hh = MCAR$ | $hh = NMAR$ | $hh = NMAR$ |
|---|---|---|---|---|
| method | $pers = MCAR$ | $pers = NMAR$ | $pers = MCAR$ | $pers = NMAR$ |
| Resp Mean | -33.4 | -20.6 | -41.2 | -26.3 |
| | ( -0.9) | ( -1.0) | ( -0.8) | ( -1.0) |
| hh resp | 210.6 | 152.5 | 303.9 | 242.0 |
| | ( -2.1) | ( -1.7) | ( -1.4) | ( -1.8) |
| donor | -47.0 | -35.8 | -55.1 | -41.0 |
| | ( -0.9) | ( -1.0) | ( -1.0) | ( -1.0) |
| class donor | -43.7 | -34.7 | -49.4 | -38.2 |
| | ( -1.0) | ( -1.0) | ( -1.0) | ( -1.0) |
| SL BLUP | -29.2 | -18.5 | -37.0 | -21.8 |
| | ( -0.9) | ( -1.0) | ( -0.8) | ( -1.0) |
| ML BLUP | 22.2 | 17.9 | 47.5 | 44.2 |
| | ( -1.4) | ( -1.1) | ( -1.2) | ( -1.2) |

Table 5: Distributional accuracy for imputing *Hourly Wage Rate* - Impute in incorrect quartile(%) (s.e.)

| Imputation<br>method | $hh = MCAR$<br>$pers = MCAR$ | $hh = MCAR$<br>$pers = NMAR$ | $hh = NMAR$<br>$pers = MCAR$ | $hh = NMAR$<br>$pers = NMAR$ |
|---|---|---|---|---|
| Resp Mean | 76.23 | 74.76 | 74.81 | 74.47 |
| | (0.09) | (0.09) | (0.09) | (0.10) |
| hh resp | 67.70 | 67.79 | 68.46 | 68.33 |
| | (0.08) | (0.09) | (0.08) | (0.08) |
| donor | 74.60 | 75.25 | 75.30 | 76.47 |
| | (0.11) | (0.12) | (0.11) | (0.08) |
| class donor | 71.92 | 72.27 | 72.65 | 73.47 |
| | (0.10) | (0.11) | (0.12) | (0.11) |
| SL BLUP | 72.39 | 72.26 | 72.63 | 74.36 |
| | (0.08) | (0.07) | (0.07) | (0.07) |
| ML BLUP | 67.44 | 65.64 | 65.65 | 67.77 |
| | (0.07) | (0.08) | (0.08) | (0.09) |