2015

Imputation of household survey data using mixed models

Luise Patricia Lago
University of Wollongong

School of Mathematics and Applied Statistics

Imputation of Household Survey Data using Mixed Models

Luise Patricia Lago

Supervisors

R.G. Clark & R.L. Chambers

This thesis is presented as required for the
Award of the Degree of
Doctor of Philosophy
of the
University of Wollongong

June 2015

Abstract

Household surveys collect information about a household and data items relating to one or more people within the household. Developing an efficient strategy for dealing with missing data is essential in the current climate of falling response rates. People within households are more likely to share characteristics than a random group of people and this homogeneity can be used when forming strategies for dealing with nonresponse. Amongst single value imputation methods, linear models and donor models are commonly used, but generally ignore relationships within households. These strategies make use of auxiliary variables available for nonrespondents to replace the missing value with a single value, for example a mean or donor value. Imputation strategies for missing items at person level will be the focus of this thesis. The goal is to make use of correlation structures within households to form improved imputed values for missing data.

Imputation models are developed and assessed using the hierarchical

structure of people within households. They are investigated for both continuous and binary missing response variables. Linear mixed imputation models, generalized linear mixed imputation models and donor imputation methods (random, within class and nearest neighbour) are investigated and compared to existing methods which do not exploit this hierarchical structure. The imputation methods are evaluated using data from two large-scale household surveys, the Household, Income and Labour Dynamics in Australia Survey (HILDA), and the British Household Panel Survey (BHPS), on a range of criteria relevant to household surveys.

For continuous variables a proposed household nearest neighbour method results in improved imputed values over other donor methods, and the success of the linear mixed model increases with the level of clustering. For binary variables the household nearest neighbour method and generalized linear mixed models both lead to improvements over standard donor and generalized linear methods.

The household imputation methods are most beneficial for improving predictive accuracy and reproducing within-household clustering in the imputed dataset. They are of some benefit for variance estimation but did not achieve much improvement over single-level methods for bias reduction. The level of improvement often depends on the assumed nonresponse mechanism, with the linear mixed model more beneficial than the household donor method

under informative nonresponse and higher levels of clustering. Otherwise, the donor household method was generally at least as good as the multilevel model and is less complex to implement.

Declaration

I, Luise Patricia Lago, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, of the School of Mathematics and Applied Statistics, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Luise Patricia Lago, June 10, 2015

Acknowledgements

This research was jointly funded by the Australian Research Council and the Australian Bureau of Statistics. I am particularly grateful to the Australian Bureau of Statistics who partly funded my scholarship. Thank you for your support, and I hope that the findings of this thesis contribute to the task of imputation for household survey data.

My primary supervisor, Robert Clark, was both patient and wise. Thank you Robert for guidance and support throughout this project and for your time spent reviewing the manuscript. Your technical knowledge, feedback and encouragement were invaluable. I am grateful to both you and my second supervisor Ray Chambers for creating an opportunity to undertake part of my studies at University of Southampton. The experience was very rewarding, and I was fortunate to discuss my research with helpful people such as Gabrielle Durrant and others at Southampton, as well as Fiona Steele and her colleagues at Bristol University. I would like to thank them for being

so generous with their time.

The management and staff at my work, the Australian Health Services Research Institute (AHSRI) have been very supportive throughout my PhD studies. Thank you to in particulary to Kathy Eagar, Rob Gordon, Janette Green, Elizabeth Cuthbert, Sam Allingham and Sonia Bird for their support. Other staff at Wollongong University provided advice and support including David Steel who generously reviewed my thesis, and also Carole Birrell, Anica Damcevski, Kerrie Gamble and Carolyn Silveri from the National Institute for Applied Statistics Research Australia.

I am grateful for being granted access to two key datasets on which I was able to investigate and evaluate the statistical methods discussed in this thesis. Firstly, this thesis 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 Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this thesis, however, are those of the author and should not be attributed to either DSS or the Melbourne Institute.

This thesis also made use of data collected in the British Household Panel Survey. The principal investigators are the University of Essex and the Institute for Social and Economic Research. The collection is sponsored by the

Economic and Social Research Council. Copyright is held by the Institute for Social and Economic Research. The survey data is distributed by the UK Data Archive, University of Essex, Colchester. The findings and views reported in this thesis are those of the author and should not be attributed to any of the preceding organisations.

The data analysis for this paper was conducted using SAS v9.2 software. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Most importantly, this thesis would not have been possible without the ongoing support and patience of my family. Thank you firstly to Ben for supporting and encouraging me and your many sacrifices to support me pursuing further study. Our children Jarred and Elise were also patient during many hours spent in the study. My parents Trevor and Kay provided encouragement and support throughout my education from school through to postgraduate study. Thanks also to both my parents and parents-in-law, Bart and Kerrie, who assisted with looking after Jarred and Elise to allow me dedicated study time.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Purpose of research 9

1.3 Scope of research 13

1.4 Thesis structure 15

2 Literature Review **19**

2.1 Notation 19

2.2 Multilevel models 21

2.2.1 Specifying the multilevel model 22

2.2.2 Variance partitioning coefficient and intra-class correlation 24

2.2.3 Modelling longitudinal data with multilevel models . . 26

2.2.4 Estimating model parameters 29

2.3 Statistical inference in the presence of missing data 32

2.3.1 The missing data pattern 32
2.3.2 The response mechanism 35
2.3.3 Methods of dealing with missing data 38
2.4 Imputation methods 40
2.4.1 Selecting auxiliary variables to define the imputation model 41
2.4.2 Linear imputation methods for household data 41
2.4.3 Donor imputation methods for household data 42
2.4.4 Multiple imputed values 47
2.4.5 Multivariate imputation 50
2.4.6 Applying multilevel methods to imputation 52
2.4.7 Imputation methods for longitudinal data 55
2.5 Evaluating Imputation Methods 56

3 Imputation of Continuous Data using Deterministic Linear Models and Linear Mixed Models 59

3.1 Introduction 59
3.2 Imputation methods 61
3.2.1 Best Linear Unbiased Predictor for single and multilevel linear model 62
3.3 Simulation study 65

3.3.1 Imputation variable from HILDA 65

3.3.2 Simulating nonresponse 67

3.3.3 Imputation methods 72

3.4 Results 77

3.4.1 Single-level and multilevel imputation compared to respondent mean 77

3.4.2 Imputation using log transform 84

3.4.3 Proportion below Federal Minimum Wage 90

3.5 Summary of Chapter 3 92

4 Imputation of Continuous Data using Stochastic Linear Models and Linear Mixed Models 97

4.1 Introduction 97

4.2 Conditional Distribution of Missing Values 99

4.3 Simulation study 104

4.4 Results 107

4.4.1 Single stochastic linear and linear mixed imputed values compared to deterministic imputed values 108

4.4.2 Multiple imputed values compared to single stochastic imputed values 113

4.4.3 Proportion of people on or below Federal Minimum Wage using deterministic and single stochastic imputation methods 118

4.5 Summary of Chapter 4 120

5 Imputation of Binary Data using Generalized Linear Models and Generalized Linear Mixed Models 123

5.1 Introduction 123

5.2 Imputation methods 125

5.2.1 Generalized Linear and Generalized Linear Mixed Model with logit link function 125

5.2.2 Generalized Linear and Generalized Linear Mixed Model with probit link function 128

5.3 Simulation study 132

5.3.1 Imputation variables from BHPS 133

5.3.2 Simulating nonresponse 134

5.4 Results 137

5.4.1 Single imputation methods 137

5.4.2 Multiple imputation methods 141

5.4.3 Multiple Imputation variance 144

5.5 Summary of Chapter 5 146

6 Imputation of Continuous or Binary Data using Donor Methods 149

6.1 Introduction 149

6.2 Donor imputation methods 152

6.2.1 Proposed imputation methods 153

6.2.2 Application of donor imputation methods in simulation study 156

6.3 Results 157

6.3.1 Nearest Significant Other compared to other donor imputed values for continuous variables 158

6.3.2 Nearest Significant Other compared to donor imputed values for binary variables 163

6.3.3 Comparison of donor methods with linear methods for continuous variables 168

6.3.4 Comparison of donor methods and stochastic generalized linear mixed methods for binary variables 174

6.4 Summary of Chapter 6 180

7 Conclusions 183

7.1 Summary of findings 183

7.2 Further research 187

List of Figures

3.1 Simulation study - estimated percentage on or below Federal Minimum wage 91

4.1 Simulation study - estimated percentage on or below Federal Minimum wage with stochastic imputed values 119

6.1 Simulation study - estimated percentage on or below Federal Minimum Wage with stochastic and donor imputed values . . 175

List of Tables

3.1 Predictive accuracy (RRMSE %) for imputing *hourly wage rate* 78

3.2 Predictive accuracy (relative bias %) for imputing *hourly wage rate* 79

3.3 Estimation accuracy for imputing *hourly wage rate* - estimated intra-class correlation 82

3.4 Estimation accuracy for imputing *hourly wage rate* - relative bias (%) of estimated variance 83

3.5 Predictive accuracy (RRMSE %) for imputing *hourly wage rate* using log transform 85

3.6 Predictive accuracy (relative bias %) for imputing *hourly wage rate* using log transform 86

3.7 Estimation accuracy for imputing *hourly wage rate* - estimated intra-class correlation with log transform 88

3.8 Estimation accuracy for imputing *hourly wage rate* - relative bias (%) of estimated variance with log transform 89

3.9 Various estimates of the percentage of adults on or below FMW 90

4.1 Efficiency for various rates of missing information and number of multiple imputed values 107

4.2 RRMSE (%) - imputation using single stochastic BLUPs compared to deterministic BLUPs 108

4.3 Relative bias (%) - imputation using single stochastic BLUPs compared to deterministic BLUPs 110

4.4 Expected value of estimated ICC (%) - imputation using sing stochastic BLUPs compared to deterministic BLUPs 111

4.5 Relative bias (%)of estimated variance - imputation using single stochastic BLUPs compared to deterministic BLUPs . . . 112

4.6 RRMSE (%) - MI compared to single stochastic imputation using SL and ML BLUPs 114

4.7 Relative bias (%) - MI compared to single stochastic imputation using SL and ML BLUPs 115

4.8 Expected value of estimated ICC - MI compared to single stochastic imputation using SL and ML BLUPs 116

4.9 Ratio of imputation variance under MI to true imputation variance under SL BLUP and ML BLUP methods 117

4.10 Various estimates of the percentage of adults on or below FMW - stochastic methods 118

5.1 BHPS simulation variables 134

5.2 RRMSE (%) - single imputation methods for Y_1 - labour and Y_2 - employed 138

5.3 Relative Bias (%) - single imputation methods for Y_1 - voting preference labour and Y_2 - employed 139

5.4 Households with same voting intention of labour (%) and same employment status (%) - single imputation methods 140

5.5 Relative Bias of population proportion (%) - single imputation methods for Y_1 - voting preference labour 141

5.6 Relative Root Mean Square Error (%) - single and multiple imputation methods for Y_1 (labour) and Y_2 (employed) 142

5.7 Relative Bias (%) - single and multiple imputation methods for Y_1 (labour) and Y_2 (employed) 143

5.8 Households with same response (%) for Y_1 (labour) and Y_2 (employed) - single and multiple imputation methods 144

5.9 Relative Bias of population proportion (%) - single and multiple imputation methods for Y_1 (labour) and Y_2 (employed) . 145

5.10 Ratio of imputation variance to true variance for population proportion (%) - for Y_1 (labour) and Y_2 (employed) 146

6.1 RRMSE (%) - imputation using nearest significant other methods compared to other donor methods 159

6.2 Relative Bias (%) - imputation using nearest significant other methods compared to other donor methods 160

6.3 Expected value of estimated ICC - imputation using nearest significant other methods compared to other donor methods . 161

6.4 Relative Bias (%) of Estimated Variance - imputation using nearest significant other methods compared to other donor methods 163

6.5 Relative Root Mean Square Error (%) - donor imputation methods for Y_1 - voting intention labour and Y_2 - employed . . 165

6.6 Relative Bias (%) - donor imputation methods for Y_1 - voting intention labour and Y_2 - employed 166

6.7 Households with same voting intention of labour (%) or same employment status (%)- donor imputation methods 167

6.8 Predictive accuracy (RRMSE %) for imputing *hourly wage rate* using linear models compared to donor methods 169

6.9 Predictive accuracy (Relative Bias %) for imputing *hourly wage rate* using linear models compared to donor methods . . 170

6.10 Estimation accuracy for imputing *hourly wage rate* - Estimated intra-class correlation using linear methods compared to donor methods 171

6.11 Estimation accuracy for imputing *hourly wage rate* - Relative Bias (%) of Estimated Variance using linear models compared to donor methods 172

6.12 Percentage of adults on or below FMW - linear methods . . . 173

6.13 Application - Percentage of adults on or below FMW - donor methods 174

6.14 Relative Root Mean Square Error (%) using linear models compared to donor methods for Y_1 - labour and Y_2 - employed 176

6.15 Relative Bias of imputed values (%) using linear models compared to donor methods for Y_1 - labour and Y_2 - employed . . 177

6.16 Relative Bias of proportion (%) for Y_1 - vote labour and Y_2 - employed using generalized linear and donor imputation methods 178

6.17 Households with same voting intention of labour (%) and employment status (%) - using generalized linear and donor imputation methods 179

7.1 Estimation accuracy for imputing *Hourly Wage Rate* - Relative Bias (%) of Estimated Mean 199

7.2 Estimation accuracy for imputing *hourly wage rate* - relative bias (%) of estimated mean using log transform 200

7.3 Rel. Bias (%) of estimated mean - MI compared to single imputation using SL and ML BLUPs 201

7.4 Rel. bias (%) of estimated mean- imputation using stochastic BLUPs compared to deterministic BLUPs 202

7.5 Rel. Bias (%) of estimated mean - imputation using NN hh compared to donor methods 203

7.6 Relative Bias of population proportion (%) - donor imputation methods for Y_1 - voting preference labour and Y_2 - employed . 204

7.7 Estimation accuracy for imputing *Hourly Wage Rate* - Rel. Bias (%) of estimated mean using linear models compared to donor methods 204

Chapter 1

Introduction

1.1 Background

Household surveys are an important source of statistical information in Australia and most nations. The Australian Bureau of Statistics (ABS) runs an extensive program of household surveys through its Population and Social Statistics Program. The current household survey program (Australian Bureau of Statistics, 2013) collects information on social topics such as labour force participation, health, family, housing, education, income, expenditure, and crime, and information on population groups such as older people, indigenous peoples, children and people with disabilities. Household surveys collect information about a household and data items relating to one or more people within the household. The structure of household survey data can

be described using a hierarchy with several levels; for example a four-level hierarchy is formed when information is available describing the area, the households selected within the area, the person or people from within the household, and different time periods at which the survey may have been repeated. Measures about the group to which an individual belongs are known as contextual variables.

Hierarchical data arise in surveys for various reasons. The focus of a survey may be the contextual variables themselves, for example an income survey may be particularly interested in measuring household income. Estimating household income requires collection of data on income from each person in the household. Data may also take a hierarchical form when the survey frame is created using a listing of groupings of people. Sample selection is then usually carried out sequentially at the different levels, for example sampling areas, then households then people. This is known as multistage sampling. Forming person-level population lists for household surveys is problematic and expensive and multistage sampling reduces this onerous task to listing all areas, the dwellings within selected areas and people within the selected dwellings. Sampling from clusters of people also generally leads to considerable cost savings not just in forming the survey frame but also in reducing travel costs where face-to-face interviews are used. In household surveys contextual variables exist at some or all of these levels. Another

reason hierarchical data arises is when clustering is of specific interest in the survey, for example the differences in crime rates between areas or change in labour force status of a person over time (longitudinal analysis). People within households and within areas are more likely to share characteristics than a random group of people. A disadvantage of multistage designs is the increased variance of estimates introduced due to this homogeneity, providing less information than a random sample from the entire population. However, within-cluster homogeneity is potentially beneficial when forming strategies for dealing with nonresponse.

Nonresponse in household surveys can occur at any level of the hierarchy - missing entire households, missing people within households, and missing data items. Longitudinal surveys may also have data missing for some time periods, however this will not be considered further, as this thesis focuses only on data collected at a single point time period. Nonresponse can occur as a result of non-contact, refusal or other reasons, such as language problems. A distinction is useful between unit nonresponse and item nonresponse. For single-level surveys nonresponding people are unit nonresponse, while nonresponse to individual data items is item nonresponse. In household surveys, unit nonresponse can be household level (when an entire household is nonresponding) or person-level (when people within the household are missing). If the household or person has partially responded this results in either house-

hold or person-level item nonresponse. In practice when there is a large amount of item nonresponse for a particular household or person, all of the information may be discarded and the person or household treated as unit nonresponse. If a household is nonresponding it will usually be excluded from the analysis dataset, and weighting methods used to compensate for the nonresponse. If a person is nonresponding within a responding household they may or may not be included on the unit record file. This has implications for calculating household totals such as household income. Person-level item nonresponse also creates problems for estimating household totals, and this is the level of nonresponse which is the focus of this thesis.

Missing data are regularly dealt with through weighting or imputation. Weighting strategies are more commonly used for dealing with unit nonresponse (household or person level), where no records exist on the analysis dataset. Weighting involves assigning each responding unit on the file one or more weights, where the weights of responding units are designed to compensate for nonrespondents and account for the use of unequal probabilities of selection in the original sample selection process. A simple weighting strategy might involve calculating weights for respondents to reflect the original sample selection probabilities, and may also involve adjusting these initial weights to meet independent population benchmark counts.

Missing data are undesirable as they can lead to bias and increased vari-

ance of point estimators (Haziza, 2009), as well as difficulty in applying standard analysis techniques, which often rely on complete data. Developing an efficient strategy for dealing with missing data is essential in the current climate of falling response rates. Strong evidence was found of increasing difficulty to make contact with households in six major U.S. household surveys (Atrostic et al., 2001), which in 1990 had initial household nonresponse rates ranging from 4.3% to 16.3% but by 1999 had worsened to between 7.5% and 28.0%. Income questions in particular have much higher item nonresponse (typically 20-40%) than non-income items (around 1-4%) (Yan, Curtin, and Jans, 2010). In the Australian context, the Household Income and Labour Dynamics of Australia Survey (HILDA) has less than 2% nonresponse for most data items (Watson, 2007), however, much higher nonresponse occurred for questions concerning income wealth and expenditure. When income components were summed for this survey, between 9% and 15% of persons had missing total income in the first five waves. For households this rose to between 22% and 29%. 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 known data (Groves and Couper, 1998) to ensure that the resulting inference has good properties (e.g. Rubin, 1996).

Imputation strategies are generally used for item nonresponse. A single value may be imputed, or more than one value to create multiple imputed

values. For an end-user of an analysis dataset a single value impute can be misleading, as the imputed value may appear to have the same certainty as respondent values. Multiple imputation strategies make more explicit the variability associated with the imputed value, and provide a means to estimate this variability. Multiple Imputation (MI) may be considered 'proper' or fractional. Proper MI involves specification of a prior distribution for the missing data values and repeated drawing of the imputed values from the posterior distribution of the missing values. Fractional imputation refers to repetitions of the imputation process which are combined in a way which depends on the process used to draw the imputed values. Both methods create more than one impute, hence multiple complete datasets can be analysed. The resulting estimators are summarised to get a mean estimate over all imputed values, and an associated variance which incorporates both the sampling and imputation variance (e.g. Rubin, 1996).

Amongst single value imputation methods, linear models and donor models are the most frequently employed. Mean imputation involves calculating the mean of the variable of interest over all respondents, then replacing missing values of the variables with this average. Mean imputation can also be carried out within classes defined by available auxiliary variables (Little, 1986b, Haziza and Beaumont, 2007) or can be derived from a regression model. Donor methods are obtained through a draw from the set of respon-

dents, which has the advantage of resulting in a continuous or binary impute depending on which is being imputed. The method of selecting a donor may be as simple as assigning a random respondent's value to a nonrespondent or may make use of auxiliary variables for which data are available on both respondent and nonrespondents to form donor classes (e.g. Kalton and Kasprzyk, 1982).

Imputation methods can be either deterministic or stochastic in nature. For example mean imputation is considered deterministic in nature as, given the sample data, the resulting impute is fixed. In contrast a random donor method is an example of a stochastic method. Donor methods can also be deterministic if there is no random mechanism in the selection of donors. Another stochastic imputation method is to take one or more draws from a fitted distribution conditional on the data.

Bankier (1999) combined editing and imputation principles in defining the imputation methodology used for the 1996 Canadian Census. Donor households were identified to create realistic imputed values within households, and these were selected to minimise the number of violated edit rules. This thesis will consider the imputation process separately from the editing strategy, and focus on predictive and estimation accuracy gains when imputing individual variables in the household setting, rather than the number of edit rules met.

While it is routine to consider whether an imputation strategy preserves univariate and multivariate population distributions (David et al. 1986 and Marker, Judkins, and Wingless 2002), in a household survey setting there are additional considerations. An important evaluation criteria specific to household surveys is the ICC. Preserving relationships with the household may be of particular importance in a household survey, for example a survey collecting data on income may aim to improve understanding of the varying income levels of people within a household. If there are strong correlations or other associations between individuals within a household there may be additional benefit in predictive accuracy in using the known values from respondents to impute nonrespondents in the same household. Clark and Steel (2002) found within-household unadjusted intra-class correlations (ICC) in the range of 0.03 (for whether a full-time student) to 0.86 (for English as a second language) with correlations typically between 0.1 and 0.3. Taking within-household correlation into account would be expected to improve both accuracy of the imputed value, and preservation of within-household ICC. If household structure is ignored in imputation then not only is potentially valuable information being disregarded, but the resulting imputed values may distort within-household patterns. Often household survey objectives include understanding relationships within households, and household-level attributes such as aggregates of person-level items. Therefore imputation

should not only accurately reproduce univariate and multivariate relationships but also within-household correlations.

Naturally, the end use of the dataset should be considered when choosing an imputation method. If the only aim is to estimate population means or totals, then intra-household correlation can safely be ignored in the imputation model, as only first order moments need be correctly specified. If the variances of estimators of means or totals are to be estimated, then the imputation model must also correctly specify the second moments of the variables requiring imputation (Haziza and Rao, 2010), although it is probably sufficient to correctly specify variances but not covariances. If within-household relationships, for example income variation within household, or even mean household income, will be considered by some analysts, then imputation should preserve these relationships.

1.2 Purpose of research

This thesis will deal specifically with imputation strategies for missing item level data. The aim is to make use of correlation structures in the data hierarchy, such as within households, to form improved imputed values for missing data.

The goal of imputation is to reduce nonresponse bias in survey estimates

and to allow analysis by complete-data methods. The purpose of this thesis is to investigate whether incorporating the multilevel structure of household survey data into the imputation model improves the quality of the resulting imputed values, for example by reducing bias and variance of estimates. In the household survey setting there is also an added dimension to the quality of the imputation method: preserving clustering within households. This will be particularly important when the imputed data are used for analysis of household attributes, or within household relationships, which are of importance in economic and social policy development.

The first model required when developing an approach to missing data is a nonresponse model. This model can be defined at household level (e.g. Groves and Couper, 1998), person level (e.g. Ezzati-Rice and Khare, 1994) or item level (e.g. Little, 1982), or a combination of these (e.g. Durrant and Steele, 2009, Wun et al., 2007). Its purpose is to describe the mechanism driving the nonresponse. The performance of any imputation method depends heavily on the mechanisms that led to the missing values (Little and Rubin, 1987, p.39). The model is required to understand what variables are correlated with response status, so that appropriate assumptions are made when forming an imputation model.

After the nonresponse model has been determined the imputation model can be developed. When developing the imputation model the ultimate use of

the data, or the analyst's model, must be considered (Schafer, 1997, p.143). This includes whether analysis will be univariate or multivariate, whether interactions will be modelled, and whether analysis will be carried out at person or household level. Household survey unit record data files are increasingly being used by large numbers of users who use a range of statistical analysis methods and so the focus of this thesis will be on imputation for general purposes, such as unit record files, and will use the hierarchical structure formed by households in developing the imputation model. A set of criteria suitable to household surveys will be used to assess the effects on estimates of proposed imputation models, including means, proportions, and variances, as well as intra-class correlations.

The estimation model, or analyst's model, is used to to combine the observed and imputed data to form survey estimates (Schafer, 1997 and Rubin, 1976). The appropriate choice of variance estimation to account for imputation variance is also part of the estimation process (Rao, 1996).

When evaluating imputation methods using simulated data a fourth model is required, the simulation model, for the values of the variable of interest. Both responding and nonresponding data can be simulated using a model. When nonresponse is simulated, factors such as the level of nonresponse, nonresponse mechanism and intra-household correlation can be modelled. Alternatively the survey respondents can be treated as the full dataset and

nonresponse simulated by applying the nonresponse process to the respondent dataset. This thesis will use real data to assess the imputation model, but also simulate higher, artificial levels of clustering to asses the ability of the imputation method in a variety of situations. Several nonresponse mechanisms will be applied at person and household level to allow evaluation of imputation methods for household data under different assumptions for the cause of nonresponse.

Mixed models are in established use for the analysis of multilevel data (Goldstein, 1995, Raudenbush and Bryk, 1992 and Raudenbush and Bryk, 1992), but are only more recently being considered for imputation of missing values in multilevel settings (Yucel, 2008, Yucel, 2008, Carpenter, Goldstein, and Kenward, 2011 and Carpenter, Goldstein, and Kenward, 2011). Imputation using mixed models has been applied, and evaluated, in datasets with reasonably large cluster sizes, for example Yucel (2008) considered imputation of children with special health care needs within states. In household surveys cluster sizes are typically very small (averaging 2-3 people per household) and ICC's may be stronger, so the use of a more complex model might be of more benefit. In Australia the distribution of household sizes is approximately as follows: 33.2% one person households, 48.8% two person households, 12.0% three person households, and 6.0% households of size 4 or more (Clark and Steel, 2007). While these methods will be of no benefit for

imputing item nonresponse for a cross-sectional survey in the 33.2% of one person households, the remaining 66.8% of households potentially have one or more respondents to draw on.

These are the fundamental questions to be addressed in this thesis: whether there is benefit in using an imputation strategy using household attributes, and in which situations (for example level of clustering or nonresponse mechanisms) is pursuing an imputation strategy using household information worthwhile?

1.3 Scope of research

The focus of this thesis is on missing item level survey data in cross-sectional household surveys. Imputation methods will be developed and assessed making use of the hierarchical structure formed by people within households. Area-level effects are a potentially useful extension as an additional level of hierarchy considered, but will not be covered. Generally speaking a higher level of ICC would be expected within households than in geographic areas, so the former is the focus of this thesis. The imputation methods will assume an all persons per household sample design.

Imputation methods will allow for missing data in the response variable, and will be investigated for both continuous and binary variables. Linear

mixed imputation models, generalized linear mixed imputation models and donor imputation methods (random, within class and nearest neighbour) will be investigated and compared. These methods will make use of information about respondents within the household in different ways in forming the imputed values for the missing data. Characteristics such as age and sex will be used as auxiliary variables as they are generally collected on the household form and are therefore widely available for all people in responding households. As this thesis includes comparisons across different imputation methods, a simple and consistent set of auxiliary variables will be chosen and applied across all models.

Producing accurate imputed values and population estimates will be central to the development of hierarchical imputation methods in this thesis. Methods for estimating variance will not be the focus but will be addressed in part using multiple imputation.

Item nonresponse can be considered missing completely at random (MCAR), missing at random (MAR) - that is dependent on some variable(s) other than the response variable, or missing not at random (MNAR) - dependent on the response variable. These terms will be further defined in Section 2.3.2. All theoretical results will be based on the assumption that data are MAR however the proposed imputation methods will be assessed via simulation studies under a range of assumptions about the mechanism for nonresponse, includ-

ing MCAR, MAR and MNAR. The nonresponse mechanism will be varied at both household and person-level.

1.4 Thesis structure

In Chapter 2 notation will be defined, followed by a review of existing literature on multilevel models, missing data frameworks and imputation methods, and their relevance to household surveys. The current literature for imputation of missing hierarchical data will be detailed, particularly item level missingness. This chapter will also include details on classifying missing data, methods of describing the missing data pattern, imputation methods, and model development.

In Chapter 3 linear mixed models for imputation will be considered for continuous variables. Single-level linear models are commonly used in imputation of hierarchical data, and the linear model will be extended to the mixed model to account for clustering present in household survey data. Model development and estimation will take place under the assumption of a continuous response variable with missing values, and a set of partially or completely observed covariates. An integral part of this thesis is the evaluation of various aspects of imputation methods including mean squared error, bias, reproduction of unit variances, and intra-household correlations. Sim-

ulated scenarios will include differing levels of within-household clustering, and various nonresponse mechanisms. The imputation methods will be applied to missing data in a simulated dataset based on real survey data, and in an application to estimating the proportion of people earning on, or below, the Federal Minimum Wage. Lastly, mean imputed values will be derived under a linear model and linear mixed model and the results compared in the household survey context.

The methods of Chapter 3 will be extended in Chapter 4 to include a stochastic component in the imputation process. Both single and multiple imputed values will be produced under the linear mixed imputation model. The stochastic component should address any issues with bias which may arise from the use of deterministic imputed values based on linear mixed models. The application to Federal Minimum Wage will be re-visited as will the consideration of imputing on the log-scale. Both Chapters 3 and 4 will use data from the HILDA survey.

Chapter 5 will investigate the use of generalized linear mixed models for imputation of missing binary data in household surveys. Generalized linear mixed models with logit and probit link functions will be compared to generalized linear models to determine whether, and when, random effects are beneficial. The comparison will include single stochastic and multiple imputed values, and the models will be applied to data from the British

Household Panel Survey (BHPS).

Donor methods are sometimes preferred over linear imputed values as the substituted data come from survey respondents, rather than a model, and hence represent actual survey realisations. Chapter 6 will look at donor methods dealing with both continuous and categorical missing data. Methods of using household characteristics to identify donors will be addressed, and existing donor imputation methods, such as random donors and within-class donors, will be compared to new methods. These household donor methods will be compared to the mixed models of Chapters 3-5 to assess their relative performance for both continuous and binary data.

The thesis will conclude with a summary of overall conclusions and implications for developing imputation methods in the household survey context. Directions for further research will then be identified.

Chapter 2

Literature Review

2.1 Notation

Generally speaking, matrices will be represented using bold upper-case (e.g. $\mathbf{A}$) and vectors will be bold lower-case (e.g. $\mathbf{y}$). To represent a column of a matrix the corresponding lower-case letter will be used in bold with the appropriate subscript to identify the particular column (e.g. $\mathbf{a_1}$ for column 1 of $\mathbf{A}$). An element of a matrix will be represented by the lower case letter in standard font with the appropriate subscripts (e.g. a_{11}). Elements of a vector will be lower-case with the appropriate subscript to identify the element (e.g. y_1 for the 1st element of the vector $\mathbf{y}$).

Assume a sample of m households is selected from a finite population U of size M. The sample has households as primary sampling units 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-class correlations are 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 $i = 1, ..., N_j$, with $n = \sum_{j=1}^{m} N_j$. 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) of size n_o and item nonrespondents $\mathbf{Y}_u$ (unobserved) of size n_u where $n = n_o + n_u$. 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 (i.e. they are in a selected household $j = 1, ..., m$) 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 $R_{ij} = 1$. For simplicity, it is assumed that there is no unit nonresponse, although this could easily be accommodated by defining the sample to consist of unit respondents only. The main difference in practice between imputing for item and unit nonresponse is that there are typically many more covariates that

can be used in models for item nonrespondents.

Let $P(.)$ denote a probability (or probability density function), and $P(.|.)$ be a conditional probability (or probability density function). $p(\mathbf{I})$ denotes the probabilities of selection for each person $ij \in U$ and hence defines the sampling mechanism. The probability distributions for $\mathbf{I}$ and $\mathbf{R}$ will be referred to as the inclusion mechanism and the response mechanism respectively (as distinct from using the term *models* which will be used to describe the probability distributions relating $\mathbf{X}$ and $\mathbf{Y}$). $P(\mathbf{I}|\mathbf{X_P}, \mathbf{Y_P})$ is the sampling mechanism (which is assumed not to depend on $\mathbf{R}$), where $\mathbf{X_P}$ and $\mathbf{Y_P}$ are the population values of X and Y. $P(\mathbf{R}|\mathbf{X}, \mathbf{Y}, \mathbf{I})$ will denote the response mechanism.

2.2 Multilevel models

Hierarchical models are generalisations of single-level models which allow parameters to vary at more than one level (Bryk and Raudenbush, 1992). In addition to being referred to as hierarchical models they are also referred to as multilevel models (e.g. Goldstein, (1995) and Feder, Nathan, and Pfeffermann, (2000)), mixed models (e.g. Breslow and Clayton, 1993) and random intercept or random coefficient models, depending on their specific form. The terms multilevel and mixed models will be used through this thesis. This sec-

tion details how multilevel models can be applied in a household survey data setting. Household survey data has an inherent hierarchical structure. The people selected in household surveys are not independent draws from an entire population, but are correlated within households (Clark and Steel, 2002). For this reason is important that statistical models looking at person-level attributes do not consider the survey participants to be independent of each other, but sharing common characteristics which can be used in modelling. Multilevel models provide a framework under which the household and person structure can be accounted for by estimating the variation in the response variable attributable to each level in the model.

2.2.1 Specifying the multilevel model

Multilevel modelling is carried out to estimate parameters for data with a hierarchical structure. This allows simultaneous modelling of the variation in the response variable at more than one level.

The general form of a two-level multilevel linear model (Goldstein, 1995, p17) with a single covariate, but allowing both the the intercept and slope to vary across clusters is:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \tag{2.1}$$

where

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

and u_{0j}, u_{1j} are random variables with $E(u_{0j}) = E(u_{1j}) = 0$, $\text{var}(u_{0j}) = \sigma^2_{u0}, \text{var}(u_{1j}) = \sigma^2_{u1}$, $\text{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$ and $\text{var}(e_{ij}) = \sigma^2_e$. The model can be equivalently expressed with a fixed part (consistent across clusters) and random part (allowed to vary across clusters) as follows:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + (\beta_{0j} + \beta_{1j} x_{ij}) + e_{ij} \tag{2.2}$$

This model differs from single-level models as there is more than one residual term, which effects the estimation procedure.

Multilevel models have been used in the analysis of clustered survey data (Carle, 2009), longitudinal data (Haynes et al., 2011), and in particular household survey data. For example Chandola et al. (2003) separated factors effecting social inequality in health at household and individual levels using a hierarchical model with households as clusters.

A standard single-level linear model relating a person-level response variable, Y_{ij} to a vector of person-level covariates $\mathbf{x}_{ij}$ and household-level covariates $\mathbf{Z}_j$ is

$$Y_{ij} = \mathbf{x}^T_{ij}\boldsymbol{\beta}_0 + \mathbf{Z}^T_j\boldsymbol{\beta}_1 + e_{ij} \tag{2.3}$$

with $e_{ij} \sim N(0, \sigma_e^2)$.

The error term e_{ij} is independent even within household, although household-level covariates are reflected in $\mathbf{Z}_j$.

In its most simple form, a single household-level random effect u_j is associated with each household:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_j^T \boldsymbol{\beta}_1 + u_j + e_{ij} \tag{2.4}$$

with $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. This is referred to as a random intercept model. Models can also be specified allowing slope parameters to vary across households however this type of model will not be investigated in this thesis. When only using person-level covariates, this model can be simplified to:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_j + e_{ij} \tag{2.5}$$

2.2.2 Variance partitioning coefficient and intra-class correlation

The intra-class correlation coefficient, also known as the intra-unit or intra-cluster correlation, is a measure of the homogeneity of a class or cluster with respect to a variable of interest. There are several definitions of the ICC, three of which will be discussed below. ICC has also been referred to

as the variance partitioning coefficient (Goldstein, 1995), which is a more general term and a more accurate description for the variance ratio in more complicated models where a simple correlation interpretation is not possible.

The most simple form of the ICC arises from a variance components model, which consists of random cluster level intercepts, but no fixed effects:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_{1j} + e_{ij} \\ e_{ij} &\sim N(0, \sigma_e^2) \\ \beta_{1j} &\sim N(0, \sigma_u^2) \end{aligned}$$

where $\text{Cov}(\beta_{1j}, e_{ij}) = 0$. In this case the ICC, ρ_{unadj}, is defined as the proportion of the total random variation in the response variable, due to the variance of the random cluster level effect (e.g. Bryk and Raudenbush, 1992, p.18):

$$\rho_{unadj} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = corr(Y_{ij}, Y_{kj}) \text{ for } i \neq k \tag{2.6}$$

The ICC is estimated by fitting the variance components model, and replacing the variance components in the ICC with their estimates from the model (e.g. West, Welch, and Galecki, 2007). The ICC in this form is constrained to be between 0 and 1 and takes high values when there is a large amount of variation between clusters relative to the variation within clusters, resulting

in a high level of homogeneity within clusters.

A second measure of intra-class correlation is the adjusted, or conditional, ICC (e.g. Raudenbush and Bryk, 1992, p.66) which measuries within-class variation after controlling for one or more covariates. The adjusted ICC (ρ_{adj}) is estimated by fitting a model with a fixed effect in addition to a random intercept, and using Equation 2.6 with the revised variance estimates. In this case the ICC represents the residual variation, after considering the contribution of the covariates, which exists within a cluster or household.

Another measure of intra-class correlation (Cochran, 1977, p.24) is a correlation coefficient between units within a cluster, and therefore may take both positive and negative values:

$$\rho_{cochran} = \frac{E[(Y_{ij} - \bar{\bar{Y}})(Y_{kj} - \bar{\bar{Y}})]}{E[(Y_{ij} - \bar{\bar{Y}})^2]} = \frac{2\sum_i \sum_{i<k} (Y_{ij} - \bar{\bar{Y}})(Y_{kj} - \bar{\bar{Y}})}{(M-1)(NM-1)S^2} \quad (2.7)$$

for clusters of equal size, where M is the cluster size, N is the total number of clusters, $S^2 = \sum_{i,j} (y_{ij} - \bar{\bar{Y}})/(NM-1)$ is the variance among all units and $\bar{\bar{Y}} = \sum y_{ij}/NM$ is the mean of the response variable per unit.

2.2.3 Modelling longitudinal data with multilevel models

Longitudinal or repeated measures data have a dependent residual structure, as it is unreasonable to assume that the residual for a person at time 1 will

be uncorrelated with their residual at time 2. This adds an extra layer of complexity to the modelling process. This layer of complexity is sometimes considered of particular interest in the survey (for example when interested in measuring variation over time) and other times is a nuisance (for example when estimating a marginal model using data from different time points). However, the issue remains that there is a correlated data structure which violates many standard statistical models used in non-longitudinal settings. Many types of models have been proposed for longitudinal data. For example, generalized linear models are used by Diggle et al. (2002) and Liang and Zeger (1986), both employing Generalized Estimating Equations to estimate the model parameters. Event history models are used to model longitudinal data (Allison, 1984) as are time series models (for example Zeger and Liang, 1986). Multilevel models have the advantage of allowing different numbers of observations observed at the lowest level, for example unobserved measurements on some occasions (Laird and Ware, 1982). This type of model is desirable when analysing unbalanced longitudinal survey data, such as survey data subject to nonresponse (either unit or item nonresponse). The following paragraphs will look specifically at multilevel models for accounting for the time-dependence among observations.

Several authors have considered a two-level model accounting for a time-dependency in longitudinal data pertaining to individuals, not allowing for

household structure. Goldstein, Healy, and Rasbash (1994) consider using a two-level model for repeated measures data, with successive observations i at level 1, clustered within individuals j at level 2, and an autoregressive term for the level 1 residuals:

$$Y_{ij} = \sum_{k} X_{ijk}\beta_k + \sum_{l=1}^{p_2} Z_{2lij}e_{2ij} + \sum_{m=1}^{p_1} Z_{1mij}e_{1mij} \tag{2.8}$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are random coefficient vectors of length p_1 and p_2 observed at levels 1 and 2 respectively, and level 1 residuals autocorrelated: $E(e_{1mij}\ e_{1mij'}) \neq 0$. One of the models for the level one residuals proposed in Goldstein, Healy, and Rasbash (1994) is a single-level AR(1) stationary model:

$$e_t = \rho e_{t-1} + v_t$$

$$\text{var}(e_t) = \sigma_e^2$$

$$\text{var}(v_t) = \sigma_v^2$$

$$E(v_t) = 0$$

which is fitted by firstly estimating ρ and σ_e^2 excluding the autoregressive component, then forming an initial estimate for the parameters associated with the time series model, and then iterating until convergence.

Three-level models for repeated measures within households over time are addressed elsewhere, e.g. Feder, Nathan, and Pfeffermann (2000) discuss

household panel surveys such as the U.S. Survey of Income and Prograamme Participation which is designed for longitudinal social and economic analysis of households and individuals, and surveys individuals every 4 months on their work history, and the three levels are formed by repeated observations (level 1) for people (level 2) within households (level 3).

Although these types of models will not be considered further in this thesis, the expansion of imputation methods from two to three levels to include area level effects or longitudinal measurement as well as household effects is an area of potential further research.

2.2.4 Estimating model parameters

Several methods are available to estimate multilevel model parameters. These include full information maximum likelihood methods such as iterative generalized least squares (IGLS) (Goldstein, 1986) or the Gauss-Newton scoring method. Also restricted or residual maximum likelihood methods using restricted iterative generalized least squares (Goldstein, 1989), expectation maximisation (EM) (Dempster, Laird, and Rubin, 1977), Fisher Scoring (Longford, 1987), and empirical Bayes estimation (Carlin and Louis, 1996) can be used.

The process of estimation by IGLS was described in Goldstein (1986)

and shown to be equivalent to maximum likelihood estimation under multivariate normality. This method generally starts with the ordinary least squares estimates of the fixed effects and iterates between estimation of the fixed coefficients and variance components. Maximum likelihood is known to produce biased estimates of the variance components (Goldstein, 1986) by ignoring the sampling variation of the fixed component, although the bias is small for large samples. Therefore Goldstein (1989) proposed a restricted IGLS and restricted maximum likelihood approach which corrected the bias issues.

The Fisher scoring algorithm (Longford, 1987) was designed to carry out maximum likelihood estimation, with the goal to converge quickly by avoiding the transformation of large matrices. The algorithm was implemented in the software VARCL.

The EM algorithm (Dempster, Laird, and Rubin, 1977) is used to estimate unknown parameters of the underlying, theorised distribution of data when some of the data are missing. The missingness problem is addressed by iteratively solving complete-data problems until convergence is reached. Both the model parameters and the non-responding data elements are considered missing and are estimated using the iterative process until convergence of the parameters to point estimates is achieved. The method can be considered informally as follows. Firstly there is an initial estimate of the model param-

eters. This allows the expectation of the missing data values to be found, using the assumed parameters to replace the unknown parameters in the model (the 'E' step). These expected values are used to create a new dataset of observed and missing data, and hence a complete-data likelihood function, which is maximised to find an improved estimate for the model parameters (the 'M'step). This parameter estimate is in turn used to re-estimate the expectation of the missing data and the process continues until convergence is achieved. The method was formalised and given its name in Dempster, Laird, and Rubin (1977), however had been in use in more specific applications prior to this time. The EM algorithm was employed by Ecochard and Clayton (1998) to estimate random effects for doubly crossed and nested hierarchical data involving repeated insemination cycles in a French donor program. The estimation of standard errors using in this approach was improved on in Clayton and Rasbash (1999) who used the stochastic data augmentation algorithm of Tanner and Wong (1987) to impute the missing data (random effects) by sampling from the distribution of the missing data conditional on the observed, and the current model parameters, then sampling new parameter values from the complete distribution to be used in the next, iterative, imputation step.

2.3 Statistical inference in the presence of missing data

This section reviews how missing data is described and measured, and its effect on statistical inference. This includes the missing data pattern, causes of the missing data, and models for the nonresponse mechanism. Strategies for dealing with missing data when making inference about the population of interest are also discussed.

2.3.1 The missing data pattern

When dealing with a dataset with missing data, a natural first step is to determine what data items have missing values, how much of each item is missing and the relative importance of the items. This will direct the level of time and resources put into dealing with the missing data. One way to summarise this information is to describe the missing data pattern, that is the pattern of missing and observed data over the full data matrix of interest. The missing data pattern in a particular sample can be expressed by listing the possible combinations of whether the variables of interest are missing or observed, and the sample frequencies associated with each.

For example, given three variables of interest, y_1, y_2 and y_3 the missing data pattern can be written as:

Missing Data Pattern	y_1	y_2	y_3	n
1	observed	observed	observed	n_A
2	observed	observed	missing	n_B
3	observed	missing	observed	n_C
4	missing	observed	observed	n_D
5	observed	missing	missing	n_E
6	missing	observed	missing	n_F
7	missing	missing	observed	n_G
8	missing	missing	missing	n_H

where n is the count of observations in the sample with each particular missing data pattern.

The missing data pattern can be described as univariate, monotone or arbitrary. Let $y_1, y_2, ...y_p$ be a set of variables of interest, where some or all of the observations for each variable may be missing. A univariate missing data pattern arises when there is missingness in only one variable. A monotone data pattern arises when the data can be ordered such that if y_k is missing, then so are $y_{k+1}, y_{k+2}, ...$, which occurs for example in drop-out for a longitudinal survey, provided dropouts do not later return to the survey. All other missing data patterns are considered arbitrary.

In the case of three variables considered above, one example of a monotone missing data pattern is when n_B, n_C, n_E and n_F are all 0 (Little and Rubin, 1987):

Missing Data Pattern	y_1	y_2	y_3	n
A	observed	observed	observed	n_A
D	missing	observed	observed	n_D
G	missing	missing	observed	n_G
H	missing	missing	missing	n_H

A monotone missing data pattern is useful for sequential imputation methods, as each variable can be imputed in sequence, starting with the variable with the least missing observations and imputing the missing cases (imputing the n_H missing cases of variable y_3 in the example above), then allowing the imputed variable to be used in the imputation model for the next imputed variable.

An arbitrary missing data pattern can make use of an initial imputation strategy to form a monotonic missing data pattern, then a sequential imputation method is used. Or an alternative imputation strategy may be developed, such as a multivariate imputation strategy which does not require a monotone missing data pattern. In a multivariate imputation strategy each missing data item is imputed using the values of all the observed data items. One method of creating a monotone missing data pattern is to delete observations which do not follow the required missingness pattern as discussed in Horton and Kleinman (2007), however this will mostly likely lead to an introduction of nonresponse bias.

2.3.2 The response mechanism

After the missing data pattern is assessed, a model for the response mechanism can be developed. Groves and Couper (1995) consider nonresponse in household interview surveys as arising from the following potential sources:

- Refusals (rf);
- Noncontacts (nc); or
- Other noninterviews (nio).

As the response rate and the nonresponse bias introduced by these groups of nonrespondents may be different, they express the mean for a variable of interest, y, in terms of each of these sources of nonresponse. For a sample of $n = r + m$ responding and missing cases, the respondent mean $\bar{y}_r$ can be expressed as a function of the full sample mean, y_n, and terms combining the response rate for each response source and the differences between the respondent mean and the mean for each nonresponse source (Groves and Couper, 1995, p.12):

$$\bar{y}_r = \bar{y}_n + \frac{m_{rf}}{n}(\bar{y}_r - \bar{y}_{rf}) + \frac{m_{nc}}{n}(\bar{y}_r - \bar{y}_{nc}) + \frac{m_{nio}}{n}(\bar{y}_r - \bar{y}_{nio}) \qquad (2.9)$$

This expression makes explicit the bias introduced by each type of nonrespondent, and that its impact depends on the nonresponse rate for that group.

Item nonresponse is a similar phenomenon, where one or more questions have nonresponse for a survey participant. Leeuw, Hox, and Huisman (2003) differentiate between three types of item nonresponse, firstly when the particular question was not responded to by the person (e.g. refused, not known, overlooked), secondly when the response was unusable (e.g. out of scope response), and finally when the useable information was lost during the survey process (e.g. data entry error). If the cause of the missingness is known, it may help develop a model for the nonresponse mechanism.

The nonresponse mechanism is often considered as a second phase of sampling, with the first phase consisting of the selected sample with known probabilities of selection, and the second phase consisting of the responding sample, with unknown probabilities of response. The probability function for the missing data mechanism is given by $f(\mathbf{R}|\mathbf{Y})$, which must be modelled. Two-phase estimation methods can then be used using both $P(I|X,Y)$ and $f(\mathbf{R}|\mathbf{Y})$.

Imputation methods are developed based on either implicit or explicit assumptions about the response mechanism. The missing data inference framework of Rubin (1987a) describes the response mechanism in distinct classes: missing at random (MAR), missing completely at random (MCAR) and not missing at random (MNAR). For a single variable Y_{ij} with observed variables x_{ij}, when the missing data mechanism is MCAR, $P(R_{ij}|Y_{ij}, I_{ij}, \mathbf{x}_{ij}) =$

$P(R_{ij}|I_{ij})$, that is the response status is independent of both the observed and unobserved data. Under MAR, $P(R_{ij}|Y_{ij}, I_{ij}, \mathbf{x}_{ij}) = P(R_{ij}|I_{ij}, \mathbf{x}_{ij})$ and the response status is random after conditioning on the observed data. When the missing data mechanism is MNAR, 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 $\mathbf{x}$ that make this assumption true. The nonresponse mechanism is considered *ignorable* when the data are MAR and the parameters in the data analysis are indepedent of the nonresponse model (Schafer, 1997). In a household survey setting both the nonresponse model and the imputation model could reasonably be expected to depend on information concerning the household or other respondents within the household. Imputation methods allowing for information about other respondents in the household are rare, however one recent example is Hayes and Watson, 2009, p.19 which describes an application where the respondent's partner's information was used in a nearest neighbour regression imputation model. The simulation study which will be described in Chapter 3 considers household-level factors in both the nonresponse and imputation models and assesses their performance relative to models without household information.

2.3.3 Methods of dealing with missing data

A broad classification for methods of dealing with missing data (Rubin, 1987a) is:

- Procedures based on completely recorded units;
- Imputation-based procedures;
- Weighting procedures;
- Model-based procedures.

The simplest method of dealing with missing data is by excluding any records with missing or partially missing data and performing analysis on completely recorded units only. This method, known as the available case method (Nordholt, 1998), can lead to significant data loss, as entire households may be excluded on the basis of a small amount of missing data, or just one missing data item.

Another approach is by imputation-based procedures, and such methods may be considered deterministic or stochastic. Deterministic methods always produce the same impute given a set of characteristics and stochastic methods have a random component. Deterministic methods often give the best results in terms of prediction of individual missing values, but typically lead to

imputed datasets with artificially low variability. Stochastic imputed values are typically more realistic in terms of their distribution and variability.

When a single impute is derived and the value is treated as observed, with standard variance estimation methods employed, the true variance can be seriously underestimated (Rao, 1996). Multiple Imputation (MI) is one way of accounting for the inflation in variance due to imputed values. Rubin (1987a, p118-119) requires a set of conditions to be met for the multiple imputed values to be considered 'proper' and the resulting inference to be valid in a Bayesian framework. Proper imputation allows the additional posterior variance due to the imputation method to be accounted for through explicit formulae (Rubin, 1988). In the case of publicly released unit record files, this approach requires the multiply imputed datasets to be available to users. Another method of producing multiple imputed values is by repeated imputation, that is repeatedly applying a stochastic imputation method such as a hot deck (Durrant, 2005).

An alternative approach to dealing with nonresponse is by weighting. Weighting is more typically used for dealing with unit nonresponse than item nonresponse. One such method is to divide the sample into adjustment cells, and to adjust the weight given to respondents by the inverse of the response rate in that cell (Little and Rubin, 1987, p55). Current developments in dealing with unit nonresponse in household surveys through weighting are

discussed in Brick (2013). One of these recent developments is the use of a multilevel logistic model of response propensity (Skinner and D'Arrigo, 2011). Inverse probability estimators based on this model were found to reduce the bias due to nonresponse when clusters are large and the intra-cluster correlation of both the survey variable and the response propensity are sufficiently high. The method worked less well for clusters containing less than 20 units but was only assessed down to clusters of size 5, so its performance on household survey data with generally smaller cluster sizes is not demonstrated.

2.4 Imputation methods

Imputation (Rubin, 1988, Little and Rubin, 1987) is a desirable approach to dealing with missing data because it can avoid the loss of partially or fully responding survey units (Little and Rubin, 1987, p.43), and results in a complete dataset, allowing standard data analysis methods to be used. This section describes imputation methods which may be applied to household data, and some issues which arise when undertaking imputation. It is assumed that a variable y contains missing values and there is available a set of fully observed covariates $\mathbf{x_{ij}}$.

2.4.1 Selecting auxiliary variables to define the imputation model

The imputation model should incorporate variables that are potentially related to the dependent variable, and also variables that are potentially related to the missingness of the dependent variable (Schafer, 1997, p.143, Sarndal and Lundstrom, 2005). Schafer (1997) also recommends that the imputation model is general enough to preserve associations among variables that may be used in subsequent analysis of the imputed dataset. The same rationale applies in household surveys to the clustering of a variable within households. In building an imputation model, the missingness of these auxiliary variables must also be taken into account. Ideally there would be no nonresponse in the auxiliary variables, but in practice this may sometimes occur. For item nonresponse, where one or a small number of items are unknown, a rich set of auxiliary variables may be available. For unit nonresponse, where a person in a responding household is a nonrespondent, the choice of auxiliary variables is more limited.

2.4.2 Linear imputation methods for household data

Linear regression models are regularly used for imputing missing continuous items (e.g. Little and Rubin, 1987, p.44) under the assumption that the re-

sponse 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 commonly made. A single-level population model for continuous Y (ignoring the second level of the hierarchy) is $Y_{ij} = \boldsymbol{X}_{ij}^T\beta + e_{ij}$ where e_{ij} are independent, identically distributed $\mathrm{N}(0, \sigma_e^2)$ random variables. The best linear unbiased predictor (BLUP) under this model using the observed data, $\boldsymbol{Y}_o$ is:

$$\hat{Y}_{\mathrm{LM},ij} = \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} \tag{2.10}$$

where $Y_{ij}^* = \hat{Y}_{LM,ij}$ when $R_{ij} = 0$ and $Y_{ij}^* = Y_{ij}$ otherwise and where $\hat{\beta}$ is the OLS estimate of the regression coefficients based on those cases for which y and x are observed. Equation (2.10) is a deterministic impute. A stochastic impute could be derived by taking a draw from the conditional distribution of the missing, given the observed data.

2.4.3 Donor imputation methods for household data

Hot deck methods impute an unobserved value using an observed value from the same survey. For this reason the term *hot deck* has more accurately been referred to as a *real-donor* impute (Laaksonen, 2002), however the term *hot deck* has been used in this thesis to be consistent with the literature. Imputation of the missing data item is carried out by careful selection of a

responding donor, whether that be a household, an individual person or a single data item. The donor is usually selected from within some category or class or by forming a predictive model. A recent review of hot deck methods (Andridge and Little, 2010) concluded that there was no consensus on the best way to apply the hot deck impute.

There are many hot deck methods available such as random selection of a donor within imputation classes, nearest neighbour imputation (Chen and Shao, 2000) and predictive mean matching (e.g. Di Zio and Guarnera, 2009 and Singh and Folsom, 2001). Penalties for repeated use of a donor can be introduced in order to maintain desirable properties of the distribution of the imputation variable. Nearest neighbour imputation involves a donor being selected by minimising a distance function related to one or more auxiliary variables, which may include geographic indictors. In the simplest case one auxiliary variable x is measured for each respondent $(x_1, y_1), ..., (x_r, y_r)$ and each nonrespondent $x_{r+1}, ..., x_n$. A missing y_k is imputed by y_l where l is the nearest neighbour of k using a simple distance function, that is l satisfies $|x_l - x_k| = \min_{1 \leq k \leq r} |x_k - x_l|$ (Chen and Shao, 2000). Predictive mean matching (Little, 1986a and Landerman, Land, and Pieper, 1997) uses the regression model to select a donor minimising the distances between predicted values from the model.

Hot deck imputation (Sande, 1983) is a stochastic method for dealing

with item nonresponse where the missing data item is replaced by a value from a respondent in the same sample. Hot deck methods are widely used in household surveys and have many useful properties including a lack of distributional assumptions, applicability to categorical data, ease of implementation, and use of actual observed data in the imputed value (Durrant, 2005).

Random donor imputation

The most straightforward donor imputation strategy is a random hot deck impute (e.g. Kalton and Kasprzyk, 1982). This method is carried out by imputing an item, y for a recipient, i, using the value from a randomly selected responding person, k, the donor: $y_i^* = y_k$ where $R_k = 1$. This method is stochastic and would be expected to do well for properties such as bias and variability (when the missingness process in MCAR) reflecting the observed response distribution. It can be used for both categorical and continuous data, and is an attractive option compared to parametric methods when the observed data are skewed or have other features that make an imputation model more complex (Durrant, 2005). However, this is a single-level approach and makes no use of auxiliary data, including household information.

Imputation of missing data items may be carried out by using values from a previous survey, or using a closely related variable. These methods

are referred to as cold deck imputation (Shao, 2000). Cold deck methods are particularly relevant for longitudinal surveys where respondents are linked across waves. Methods to impute using data from previous waves have been investigated for example the Little and Su (1989) and random carry-over method (Williams. and Bailey, 1996).

Class donor imputation

When information on covariates is available for both respondents and nonrespondents this information can be used to reduce nonresponse bias. For example Little and Rubin (1987) describe hot deck imputation within adjustment cells where the sample is divided into distinct imputation cells, within which response is assumed ignorable, and a missing value in that cell is replaced by a respondent from the same cell. Since the hot deck imputed values are actual values from the sample, the distribution of the imputed values is not distorted like a mean impute (Little and Rubin, 1987, p.64).

Imputation classes offer an improvement on the random donor as auxiliary information can be incorporated into the selection of an appropriate donor by identifying observed variables related to the imputation variables (see also Kalton and Kasprzyk, 1982). Imputation classes are defined by observed, usually categorial variables, or alternatively by grouping continuous variables, to create pools of potential donors which match the person with

item nonresponse on a set of observed variables. For example the value of y from a randomly selected donor may be used to impute a nonrespondent (the recipient) who is the same sex and age group as the donor. This method also has a stochastic component and should improve predictive accuracy over the random donor assuming an appropriate selection of variables to define the imputation classes. This method also fails to account for any clustering which may be present within households, with imputation classes generally using only characteristics of the nonrespondent. Attributes of the household, or any responding household members are not typically used to define imputation classes for item nonresponse at person level.

Nearest neighbour imputation

Nearest neighbour imputation is carried out by selecting a donor to minimise a measure of distance between the donor and the recipient. The measure of distance is defined using one or more auxiliary variables. Although it is a deterministic method, the nearest neighbour impute was shown by Chen and Shao (2000) to lead to low bias of estimated means, totals, quantiles and distributions, and also good variance properties. While the sample mean was shown to be asymptotically unbiased, the bias was of order r^{-1} where r is the number of respondents. The nearest neighbour imputation method can be implemented using a simple distance measure for a single auxiliary

variable as follows. For missing y_i the impute is derived by considering an auxiliary variable x_i and determining the nearest neighbour to person i. The nearest neighbour is determined by considering the set of bivariate pairs $(y_1, x_1), ..., (y_r, x_r)$ where all values of x and y are assumed to be completely observed. The donor value for missing y for person i is y_k where k is the person with $min|x_i - x_k| : i, k \in \mathcal{R}$ where $\mathcal{R} = \{1, ..., r\}$ is the set of responding units. If x is continuous the imputed person is simply the responding person with the closest value of x. When a categorical variable is used the method may result in more than one potential donor with the minimum distance, and a donor is selected from the potential set (e.g. by random selection). When more than one auxiliary variable is used to identify a nearest neighbour the distance measure must be multivariate. The Mahalanobis distance can be used for this purpose: $(\mathbf{x}_i - \mathbf{x}_j)^T \hat{\mathbf{V}}_\mathbf{i}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$ where $\hat{\mathbf{V}}_\mathbf{i}$ is the estimated variance-covariance matrix of $\mathbf{x_i}$.

2.4.4 Multiple imputed values

Unlike single imputation methods, multiple imputation replaces each missing value by two or more values to reflect uncertainty in the imputed value. Multiple imputation has important advantages over single-value imputation methods. The variance due to the unknown missing data values can be in-

corporated into variance estimates for parameter estimates. Also the impute itself may be less likely to be considered a known data value by end-users, as its uncertainty is explicitly shown. Disadvantages of multiple imputation include the extra time required to develop and apply an appropriate method, the increased variables required on the datafile for the m imputed values, and the added complexity introduced by requiring repeated analysis and combination of these analyses. This may be a considerable practical issue if the processing time is long, or the time allowed for survey analysis is very short.

A set of m datasets can be created by replacing the missing values in the dataset with each of the multiply imputed sets of data values. Rubin (1996) describes how the resulting estimates from these datasets can be combined for inference purposes. Suppose that a scalar quantity of inference, Q is estimated using each complete dataset, resulting in estimates $\hat{Q}_j$ from each dataset $j = 1, 2, ..., m$. The overall estimate of Q is given by the average of the m estimates:

$$\hat{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}_j \tag{2.11}$$

Let $\hat{U}_j$ be the estimated variance associated with the estimate $\hat{Q}_j$, estimated by treating imputed values as actual observations using complete data methods. The within and between imputation variance must then be

calculated:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j \tag{2.12}$$

$$\hat{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \hat{Q})^2 \tag{2.13}$$

These are then combined to determine the variance of the estimate over multiple imputed values:

$$\hat{\text{var}}(\hat{Q}) = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{2.14}$$

$$= \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j + \left(1 + \frac{1}{m}\right)\frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \hat{Q}\right)^2 \tag{2.15}$$

The first term estimates the variance under complete response while the second term estimates the variance due to imputation uncertainty.

The imputed values are stored in a separate dataset which has m columns reflecting the repeated imputed values, and r rows where r is the number of missing values in the survey dataset. Early literature suggested as little as 3-5 imputed values may be sufficient (Rubin, 1987a) however more recent literature (Bodner (2008), White, Royston, and Wood (2009)) which considers not just efficiency, but also quantities such as p-values and confidence intervals, recommends the number of imputations being similar to the percentage of cases that are incomplete, e.g. 17% of cases have missing data on one or more variables in the data analysis model would then require 20 imputed values.

Fractional imputation is a related method which was originally proposed in Kalton and Kish (1984) and involved repeated hot deck imputed values with fractional weights for each of the imputed values. The aim of fractional imputation is to improve the efficiency of the point estimator. Repeated imputation simplifies variance estimation and reduces the random component of the variance arising from imputation. This method can preserve the distribution of the variable being imputed, makes no distributional assumptions and imputes actual observed values. Fractional imputation has a major practical advantage in that the repeated imputations need not be stored, just the imputation weights reflecting the number of times a donor has been used for imputation. However, fractional imputed values are not proper in the (Bayesian) sense of Rubin, and the theoretical justification for their use in variance estimation is much less clear. Bjornstad (2007) addresses non-Bayesian imputation, typically employed in national statistical institutes, and describes alternative ways of combining multiple imputed values under certain response mechanisms and hot-deck type imputed values.

2.4.5 Multivariate imputation

The previous sections have addressed one or multiple imputed values for a single variable requiring imputation. One approach to imputing multiple

variables is imputation by chained equations (ICE) or multiple imputation by chained equations (MICE) (Buuren and Groothuis-Oudshoorn, 2011). It involves iteratively fitting a series of univariate models, rather than specifying a full multivariate model. However, recent research such as Robbins, Ghosh, and Habiger (2013) and Borgoni and Berrington (2013) has focussed on developing a multivariate imputation strategy which simultaneously addresses missingness in more than one variable. Robbins, Ghosh, and Habiger (2013) link imputation variables through a multivariate Gaussian distribution (after appropriate transformations) and use a regression approach used to select flexible conditional models, termed iterative sequential regression. Little and Schluchter (1985) address the issue of jointly imputing categorical and continuous variables for either estimation or imputation using maximum likelihood estimation and the EM algorithm. Borgoni and Berrington (2013) use a tree-based approach to multivariate imputation. Multivariate imputation methods are useful for preserving relationships across variables within a survey. However, the focus of this thesis is on household surveys, and in particular on preserving the relationships of a single variable within a household.

2.4.6 Applying multilevel methods to imputation

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.

Multilevel MI has been implemented in the statistical package REALCOM (Carpenter, Goldstein, and Kenward, 2011) which allows multilevel MI of continuous, ordinal and unordered categorical data, and allows missing data at level 1 or level 2. Examples of implementation has focussed on students within classes and longitudinal data (Goldstein et al., 2009).

Multivariate, multilevel methods can address both within-household and across variable relationships simultaneously. This area is just beginning to be explored for example by Yucel (2008), who extends multilevel imputation models to multivariate applications with missing data at any level of the hierarchy. This was further addressed in Yucel (2011) which developed algorithms for multivariate multiple imputation using MCMC with flexible covariance structure. Multivariate imputation models typically require a rich set of covariates to ensure that the imputation model performs well across a

large number of variables.

Imputation was considered in cluster sampling (Shao, 2007), using a linear mixed model with a cluster-level random intercept. The probability of nonresponse was allowed to depend on the unobserved random intercepts, so that response was non-ignorable, and it was assumed that there is at least one respondent per cluster. Under this model, the respondent mean from the same cluster was shown to be an unbiased impute for non-responding units. This approach may be unstable when there are few respondents in some clusters, as happens when clusters are households. An alternative also proposed by Shao (2007) is to use the respondent mean for all clusters with the same size and response rate as the cluster requiring an impute. A jack-knife variance estimator was recommended for imputation variance. In this thesis, ignorable nonresponse conditional on covariates will be considered, which is more restrictive than Shao (2007), although the impact of non-ignorable nonresponse will also be considered in the simulation study. Both person-level and household-level covariates will be able to used in imputation methods considered, as will the effect of imputation on preserving relationships within households. The methods of Shao (2007) would lead to the intra-household correlation being too high in the imputed dataset, which may be a concern in some household surveys.

Yuan and Little (2007) developed Bayesian multiple imputation methods

for two-stage sampling with item nonresponse. A linear mixed model was used for the variable of interest, Y, conditional on covariates, with a random intercept for each cluster. This was supplemented by a similar two-level random intercept model for response propensity, $P(R = 1|X, Y)$, denoted Z. The possibility of ignorable nonresponse is discussed, but the key model assumes non-ignorable nonresponse, because Y and Z are independent given the random intercepts, and the random intercepts are also independent. They also consider a general functional form of $E[Y|z]$ which can be estimated by spline or kernel regression. To avoid the consequent "curse of dimensionality" when there are many covariates, they use the modelled response propensity as a covariate in the model for Y, which can then include fewer other covariates. The main example used was the US National Health and Nutrition Examination Survey, where clusters were counties, and units were people.

The methods of Yuan and Little (2007) may be less applicable to households as clusters, because the very small number of units per cluster (only 1, 2 adults in more than half of households) may mean that the parameters of the two-level response propensity model can only be imprecisely estimated. In addition, it is not clear that the additional complexity of a response propensity model is necessary unless there are many potential covariates, because response is still assumed to be ignorable. This thesis will focus on the case of people within households, using a simpler model, but giving more attention

to the preservation of within-household structures, which is only important when households are clusters.

2.4.7 Imputation methods for longitudinal data

Like household survey data, longitudinal data has a natural hierarchy which may be incorporated into an imputation model. Imputation models can then incorporate potentially rich sources of auxiliary information on the item-nonrespondent from previous (or future) waves. Methods such as carry forward (or back) (Williams. and Bailey, 1996) or a multivariate model across waves (Little and Su, 1989) make use of data from other waves for imputation through a single-level modelling approach.

Pfeffermann (1988) developed *augmented regression predictors* by making 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 linear mixed 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. A simulation study using data generated for households of size two or three with an assumed ICC of 0.4 across four time points found some benefits in predictive accuracy in incorporating

the household information but did not explore the benefits of these models in regards to ability to reproduce ICC, and the results are based on artificial data. An empirical study looked at imputing number of hours worked during the week preceding the interview. The authors evaluated this method on labour force data over four waves, for 567 people within 475 households. To overcome problems with convergence in model estimation and negative variance estimates under IGLS, model parameters were estimated using state space methods. Due to most of the households in this study having only one person, most of the advantage of the hierarchical modelling resulted from the clustering of observations over time. The advantage of clustering within households was thus unable to be explored using empirical data. The reasons suggested were the fit of the model no longer being perfect, small household sizes (most with just one person) and smaller sample size for parameter estimation.

2.5 Evaluating Imputation Methods

An important part of the imputation process is evaluation of the imputation strategy. Ideally the analysis model is pre-determined and the imputation method for the missing variable then can be evaluated by its ability to reproduce any complete data analysis. For example missing values can be gener-

ated in a complete dataset and alternative imputation strategies compared. In Chambers (2001) this is termed *preservation of analysis*. Laaksonen (2005) also described an 'Integrated Modelling Approach to Imputation' which includes a first step of selecting a training dataset and auxiliary variables for evaluation, which is carried out prior to the construction of the imputation model. 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 described 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. Two of Chambers' criteria were used to evaluate the imputation models in this thesis; *Predictive accuracy* refers to the performance of the imputation model in reproducing the true values; *Estimation accuracy* considers the performance of the imputation methods in reproducing first and second order moments 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) generated nonresponse for households of size two or three, under MCAR, MAR and MNAR models with 20%

nonresponse. Each imputation method was applied to 100 simulated samples. Relative Root Mean Square Error (RRMSE) and Relative Bias was used to compare the predictive accuracy of imputed values with known values for various imputation methods. The RRMSE of the imputed values over K replicates of the non-response mechanism can be calculated as follows:

$$\begin{aligned} \text{RRMSE}_{av} &= \frac{1}{K}\sum_{k=1}^{K}\text{RRMSE}_k \\ &= \frac{1}{K}\sum_{k=1}^{K}\left\{\sqrt{\frac{\sum_{ij\epsilon S_k}\left(y^*_{ij,k}-y_{ij}\right)^2}{\sum_{ij\epsilon S_k}\left(1-R_{ij,k}\right)}}\Bigg/\frac{\sum_{ij\epsilon S_k}\left(1-R_{ij,k}\right)y_{ij}}{\sum_{ij\epsilon S_k}\left(1-R_{ij,k}\right)}\right\} \end{aligned} \tag{2.16}$$

where S_k is the k-th replicate sample. Relative bias of the imputed values is calculated for each replicate and averaged over the K replicates:

$$\text{RBias}_{av} = \frac{1}{K}\sum_{k=1}^{K}\text{RBias}_k = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{ij\epsilon S_k}\left(y^*_{ij,k}-y_{ij}\right)}{\sum_{ij\epsilon S_k}y_{ij}(1-R_{ij})} \tag{2.17}$$

Note that (2.16) and (2.17) refer to the properties of a set of imputed values, which is different from the usual usage where RRMSE and bias refer to an estimator, and therefore the RRMSE given above is only calculated for imputed values.

Chapter 3

Imputation of Continuous Data using Deterministic Linear Models and Linear Mixed Models

3.1 Introduction

Mixed models are in established use for the analysis of multilevel data (Goldstein, 1995, Raudenbush and Bryk, 1992), but are only more recently being considered for imputation of missing values in multilevel settings (Yucel,

2008, Carpenter, Goldstein, and Kenward, 2011). Imputation using mixed models has been applied, and evaluated, in datasets with reasonably large cluster sizes, for example Yucel (2008) considered imputation of whether mental health care was needed but not received, for children with special health care needs, with individual children clustered within states. However, variables of interest within geographic clusters may have weaker intra-cluster correlations than variables measured in household surveys.

This chapter will focus on the use of linear multilevel imputation models in household surveys where more than one person in the household is selected. These surveys are of particular interest as they raise the possibility of making use of one or more respondents within a household to impute its nonrespondents. Imputation methods will be considered for an outcome variable of interest, making use of auxiliary variables available for both respondents and nonrespondents in the household.

The aim is to investigate imputation in a 2-level linear mixed model for people within households and compare to single-level approaches. The simplest single-level model will contain only information about the item nonrespondent, while another will incorporate information from an item-respondent in the household as a covariate. The latter model is a simpler alternative to a multilevel model while still incorporating information from another member, or members, of the household, without explicitly modelling

the household effect. The comparison will be carried out with varying levels of ICC and under different nonresponse mechanisms to assess when the mixed imputation model is most beneficial, for example: does the improvement in imputed values from using a multilevel imputation model increase as the ICC increases? Is a single-level model adequate when nonresponse is Missing Completely At Random (MCAR)? Are there any benefits in using a multilevel imputation model over a single-level model incorporating information about the household? Methods resulting in a single impute will be considered for a continuous outcome variable to assess the relative merit of multilevel compared to single-level imputation models in the household setting.

Section 3.2 details the imputation models considered, using the methods described in Section 2.5. Section 3.3 describes a simulation of imputation under informative and non-informative missingness, and Section 3.4 contains results. Section 3.5 will draw conclusions and discuss areas for further investigation.

3.2 Imputation methods

This section describes the BLUP of missing Y_{ij} under a single and multilevel model given completely observed $\mathbf{x}_{ij}$.

3.2.1 Best Linear Unbiased Predictor for single and multilevel linear model

When a variable of interest is likely to be correlated within households linear mixed models may be used, which incorporate this correlation. A mixed model (Goldstein, 1995, West, Welch, and Galecki, 2007) treats regression coefficients as random variables with a different realisation for each household.

The two-level linear mixed model with a single covariate x_{ij} is $Y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j} x_{ij}) + e_{ij}$. 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 $var(u_{1j}) = \sigma^2_{u1}$ and $cov(u_{0j}, u_{1j}) = \sigma_{u01}$, $u_{0j} \sim N(0, \sigma^2_{u0})$, $u_{1j} \sim N(0, \sigma^2_{u1})$, $cov(u_{0j}, e_{ij}) = cov(u_{1j}, e_{ij}) = 0$.

Households only contain a small number of people (often just one), so a special case, the random intercept model, is typically used otherwise the additional parameters associated with random slopes are likely to lead to convergence issues due to instability. This restricts the random component to the intercept term only: $Y_{ij} = \beta_0 + \beta_1 \mathbf{x}_{ij} + u_{0j} + e_{ij}$ where $\mathbf{x}_{ij}$ now represents a vector of covariates. Henceforth u_{0j} will be written as u_j for simplicity, $\boldsymbol{\beta}$ for the vector of regression coefficients associated with the fixed part of the model, and u_j and e_{ij} for household and person-level residuals respectively.

Correlation of a continuous variable within households is measured by the ICC, defined as the proportion of total variation due to clustering within households, $ICC = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ (West, Welch, and Galecki, 2007, p.98). 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.

Under a linear mixed 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, $\hat{Y}_{\text{LM},ij} = (\mathbf{x}_{ij}^T\mathbf{x}_{ij})^{-1}(\mathbf{x}_{ij}^T Y_{ij})$, (which has no across unit correlation) plus a term incorporating the within-household covariance (for derivation see Appendix A):

$$\hat{Y}_{\text{LMM},ij} = \hat{Y}_{\text{LM},ij} + C(Y_{ij}, \mathbf{y}_o)\mathbf{V}_o^{-1}\{\mathbf{y}_o - \mathbf{x}_o(\mathbf{x}_o^T\mathbf{V}_o^{-1}\mathbf{x}_o)^{-1}(\mathbf{x}_o^T V_o^{-1}\mathbf{y}_o)\} \quad (3.1)$$

where $C(\mathbf{y}_o, Y_{ij})$ is a vector of covariances between the observed $\mathbf{y}_o$ and the missing value Y_{ij}, $\mathbf{V}_o$ is a block diagonal matrix with blocks $\mathbf{V}_{o,j} = \sigma^2((1-\rho)\mathbf{I}_{n_{o,j}} + \rho\mathbf{1}_{n_{o,j}}\mathbf{1}_{n_{o,j}}^T)$ 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 $n_{o,j}$.

There is no covariance between people in different households, so this can be simplified to:

$$\hat{Y}_{\text{LMM},ij} = \hat{Y}_{\text{LM},ij} + C(Y_{ij}, \mathbf{y}_{o,j})\mathbf{V}_{o,j}^{-1}\{\mathbf{y}_{o,j} - \mathbf{x}_{o,j}(\mathbf{x}_{o,j}^T\mathbf{V}_{o,j}^{-1}\mathbf{x}_{o,j})^{-1}(\mathbf{x}_{o,j}^T V_{o,j}^{-1}\mathbf{y}_{o,j})\} \tag{3.2}$$

Barroso, Bussab, and Knott (1998) derived a general form of Equation (3.2), 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 thesis specifically looks at cross-sectional household survey data.

In a household survey with nonresponse, the V_{0j} depend on the unknown variance parameters σ_u^2 and σ_e^2. The predictions resulting from substituting estimates for these variance parameters is known as the empirical BLUP (Barroso, Bussab, and Knott, 1998). Several estimators of σ_u^2 and σ_e^2 are available, including Maximum Likelihood, Restricted Estimation by Maximum Likelihood (Patterson and Thompson, 1971), and Minimum Variance Quadratic Unbiased Estimation (Searle, Casella, and McCulloch, 1992). The first two of these methods can be implemented using iterative techniques while the latter provides a non-iterative alternative with reduced processing time and not requiring normality (Wang, Xie, and Fisher, 2012).

3.3 Simulation study

3.3.1 Imputation variable from HILDA

A simulation study was carried out by applying a set of imputation models to a continuous variable from 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 item nonresponse. Hourly wage rate was selected for the simulation study, over other income variables such as total wages, as it was expected to be ,ore highly correlated within households. For example a negative within household correlation ($\rho = -0.04$) was found for hours worked in two-adult households (Gregg and Wadsworth, 1996) which may affect total wages. The same study found education ($\rho = 0.43$) and in particular age ($\rho = 0.91$) are highly correlated within households and these factors are likely to be reflected in the hourly wage rate.

The sample was subset 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 imputed values compared to known values. As the imputation method is designed to make use of the responses from one or more people within the household, the sample was

restricted to households with two respondents to hourly wage rate. The restriction to two-person households is made throughout the thesis. This resulted in a sample of 2,392 persons from 1,199 households, representing approximately 50% of the responding sample.

The unadjusted ICC for hourly wage rate in the sample was 0.194, which is equivalent to 19.4% of the total variance being explained by the household level for the mean model. This was estimated by fitting an intercept only model. To assess the multilevel (ML) BLUP against the single level (SL) BLUP under different levels of clustering, the pairing of some people within households was artificially adjusted to create households where the hourly wages were more similar. This was done by generating bivariate normal random variables with unit variance and different levels of clustering, ρ, within each household as follows: $Y_{ij} = Z_{ij}\sqrt{1-p^2} + Z_j p$ where $Z_{ij}, Z_j \sim N(0,1)$. The constant p was empirically selected to achieve a moderate ICC and a high ICC. Each value of Y_{ij} was converted to a rank using $R_{ij} = 1+\text{floor}(n\Phi(Y_{ij}))$. The new hourly wage rate and covariate set were then taken from the equivalently ranked value of hourly wage rate in the set of all hourly wage rates, artificially placing people of more similar wage rate in the same household. Three scenarios were selected, the true ICC of $\rho = 19.4\%$, $\rho = 50.0\%$ and $\rho = 85.0\%$, and these were used to assess the imputation methods under a low, moderate and high ICC. These will be referred to in the tables as

$\rho = 20\%$, $\rho = 50\%$ and $\rho = 85\%$ for ease of reference.

3.3.2 Simulating nonresponse

The fully observed component of the sample was used to generate $K = 250$ simulated samples with item nonresponse, to isolate the impact of the item nonresponse mechanism and imputation method as distinct from population or sample variation. Approximately half of households were designated to have item nonresponse and one of the two people within each nonresponding household was selected to be an item nonrespondent according to the different response models described below. The resulting item response rate was approximately 75% under each scenario. Five alternative models were used to generate nonresponse. The first has data Missing Completely At Random (MCAR), the second Missing at Random (MAR) and the other three Missing Not At Random (MNAR). The auxiliary variables used to define the MAR nonresponse mechanism are described below.

Let p_{1j} and p_{2j} be the probabilities of response for person 1 and person 2 in household j respectively. The following notation will be used to represent

the probability of each possible household response pattern:

$$q_{j(0,0)} = \text{P(both person 1 and person 2 nonrespondent)}$$

$$q_{j(1,0)} = \text{P(person 1 respondent and person 2 nonrespondent)}$$

$$q_{j(0,1)} = \text{P(person 1 nonrespondent and person 2 respondent)}$$

$$q_{j(1,1)} = \text{P(both person 1 and person 2 respondent)}$$

The probabilities of each possible household response pattern are $q_{j(0,0)} = 0$; $q_{j(1,0)} + q_{j(1,1)} = p_{1j}$; $q_{j(0,1)} + q_{j(1,1)} = p_{2j}$; and $q_{j(1,1)} = 1 - q_{j(0,0)} - q_{j(1,0)} - q_{j(0,1)}$ which can be manipulated to give:

$$\begin{aligned} q_{j(0,0)} &= 0 \\ q_{j(1,0)} &= 1 - p_{2j} \\ q_{j(0,1)} &= 1 - p_{1j} \\ q_{j(1,1)} &= 1 - q_{j(1,0)} - q_{j(0,1)} \\ &= p_{1j} + p_{2j} - 1 \end{aligned}$$

The probabilities will now be specified for each nonresponse scenario. In all five scenarios, the average of p_{ij} was exactly or approximately 0.75, so that approximately 75% of people were item respondents, leading to 50% of households having full response and 50% of households having one item nonrespondent. No households were simulated with no respondents, as the focus is on imputing item nonresponse using a responding household member,

so $q_{j(0,0)}$ will not be specified for all the models.

Each scenario below describes the mechanism for whether a household is fully responding or has item nonresponse, and the mechanism for selecting the nonresponding person in households that are not fully responding.

1. Households MCAR and persons MCAR: In partially responding households, one person was randomly chosen to be the full respondent, the other to have item nonresponse. Partial response probabilities are $p_{ij} = 0.75$ for all people, so that $q_{j(1,0)} = q_{j(0,1)} = 0.25$ and $q_{j(1,1)} = 0.5$ for all households.

2. Households MCAR and persons MAR: A MAR nonresponse mechanism was created by letting the odds ratio for item nonresponse be approximately 2.2 ($exp(0.8)$) for males and for those aged under 30. The probability of response at person level was specified using a logistic model as follows:

$$p_{ij} = \frac{\exp(\beta_0 - 0.8X_{1,ij} - 0.8X_{2,ij})}{1 + \exp(\beta_0 - 0.8X_{1,ij} - 0.8X_{2,ij})}$$

where $X_{1,ij}$ and $X_{2,ij}$ are indicator variables for male and age < 30 respectively. β_0 was set such that approximately 75% of people respond to the income questions, in this and subsequent models. The odds ratio of 2.2 results in approximately 33% probability of response for males under 30 years, compared to an overall response rate of 75%. The probability of each house-

hold response pattern was then:

$$q_{j(1,0)} = 1 - \frac{\exp(\beta_0 - 0.8X_{1,2j} - 0.8X_{2,2j})}{1 + \exp(\beta_0 - 0.8X_{1,2j} - 0.8X_{2,2j})}$$
$$q_{j(0,1)} = 1 - \frac{\exp(\beta_0 - 0.8X_{1,1j} - 0.8X_{2,1j})}{1 + \exp(\beta_0 - 0.8X_{1,1j} - 0.8X_{2,1j})}$$
$$q_{j(1,1)} = \frac{\exp(\beta_0 - 0.8X_{1,1j} - 0.8X_{2,1j})}{1 + \exp(\beta_0 - 0.8X_{1,1j} - 0.8X_{2,1j})} + \frac{\exp(\beta_0 - 0.8X_{1,2j} - 0.8X_{2,2j})}{1 + \exp(\beta_0 - 0.8X_{1,2j} - 0.8X_{2,2j})} - 1$$

3. Households MCAR and persons MNAR: The probability of being a respondent was dependent on Y_{ij}, where an increase in hourly wage of one dollar was associated with a 1% decrease in the odds of response. See for example Lillard and Smith (1986) who found nonresponse propensity for earnings depended on income level and was lower in the tails of the distribution, but most substantially so in high-earning income brackets. Hence the nonresponse process is determined by:

$$p_{ij} = \frac{\exp(\beta_0 - 0.01Y_{ij})}{1 + \exp(\beta_0 - 0.01Y_{ij})}$$

The probabilities associated with each household response pattern under this scenario were:

$$q_{j(1,0)} = 1 - \frac{\exp(\beta_0 - 0.01Y_{2j})}{1 + \exp(\beta_0 - 0.01Y_{2j})}$$
$$q_{j(0,1)} = 1 - \frac{\exp(\beta_0 - 0.01Y_{1j})}{1 + \exp(\beta_0 - 0.01Y_{1j})}$$
$$q_{j(1,1)} = \frac{\exp(\beta_0 - 0.01Y_{1j})}{1 + \exp(\beta_0 - 0.01Y_{1j})} + \frac{\exp(\beta_0 - 0.01Y_{2j})}{1 + \exp(\beta_0 - 0.01Y_{2j})} - 1$$

4. Households MNAR and persons MCAR: Households are partially or fully responding, with the probability of the household falling in the first category decreasing by 1% with each dollar increase in average household hourly wage rate. Within partially responding households, one person was randomly selected to be an item respondent, the other had item nonresponse. For this situation we used:

$$p_{ij} = \frac{\exp(\beta_0 - 0.01\bar{Y}_j)}{1 + \exp(\beta_0 - 0.01\bar{Y}_j)}$$

where $\bar{Y}_j = (Y_{1j} + Y_{2j})/2$. The probabilities associated with each household response pattern under this scenario were:

$$q_{j(1,0)} = 1 - \frac{\exp(\beta_0 - 0.01\bar{Y}_j)}{1 + \exp(\beta_0 - 0.01\bar{Y}_j)}$$

$$q_{j(0,1)} = q_{j(1,0)}$$

$$\begin{aligned} q_{j(1,1)} &= p_{1j} + p_{2j} - 1 \\ &= 2\frac{\exp(\beta_0 - 0.01\bar{Y}_j)}{1 + \exp(\beta_0 - 0.01\bar{Y}_j)} - 1 \end{aligned}$$

5. Households MNAR and persons MNAR: Finally, households and persons were assigned to be partially or fully responding, both MNAR. The probability of the household falling in the first category decreased by 1% with each dollar increase in average household hourly wage rate, and each persons probability of response also decreased by 1% with each dollar increase in

average hourly wage rate. This implies that person probabilities of response depend on household mean income:

$$p_{ij} = \frac{\exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{ij})}{1 + \exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{ij})}$$

The probabilities associated with each household response pattern under this last scenario were:

$$q_{j(0,0)} = 0$$

$$q_{j(1,0)} = 1 - \frac{\exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{2j})}{1 + \exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{2j})}$$

$$q_{j(0,1)} = 1 - \frac{\exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{1j})}{1 + \exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{1j})}$$

$$q_{j(1,1)} = \frac{\exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{2j})}{1 + \exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{2j})} + \frac{\exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{1j})}{1 + \exp(\beta_0 - 0.01\bar{Y}_j - 0.01Y_{1j})} - 1$$

3.3.3 Imputation methods

Four 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:

- *Respondent Mean:* $\hat{Y}_{ij}$ = mean of Y over all fully responding people in the sample.

- *Deterministic Single-level BLUP:* empirical BLUP for single-level linear model, as in Equation (2.10) in Section 2.4.2 with age and sex as covariates (notated 'SL' in tables).

- *Deterministic Single-level BLUP:* empirical BLUP for single-level linear model, as in Equation (2.10) in Section 2.4.2 with age and sex, plus co-householder response as covariates (notated 'SL+' in tables).

- *Deterministic Multilevel BLUP:* empirical BLUP for linear mixed model, as in Equation (3.2) with age and sex as covariates (notated 'ML' in tables).

The BLUP imputation models used age group by sex as explanatory variables as these would be readily available on most household survey forms, and are unlikely to themselves be subject to high levels of nonresponse. They are therefore likely to be available even for people in households with unit or item nonresponse, where there may be a large number of other data items with nonresponse. The age groups were 16-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 and 60+. These same variables were used to define all imputation classes and imputation models throughout the thesis to ensure comparisons of imputation methods were not influenced by differences in auxiliary variables. Additional variables such as education and employment status would potentially make more rich imputation models, but could have made the simulation study inconsistent across imputation methods, for example if imputation classes for donor methods needed to be collapsed. Hence a simple set of covariates were selected which could be applied consistently

throughout the thesis.

Both the linear and linear mixed imputation methods were calculated initially using untransformed data. The simulation was also carried out with a log transform for hourly wage rate, and each of the evaluation criteria calculated. A log transform of the outcome variable was performed prior to imputation. A linear model on the log-transformed data results in predictions based on the following model: $log(Y_{ij}) = \boldsymbol{\beta}^T \mathbf{x}_{ij} + e_{ij}$ where $e_{ij} \sim N(0, \sigma^2)$. Back-transformation to the original scale results in imputed values with expected values estimated by $E[\hat{Y}_{ij}] = \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + \frac{s^2}{2})$ (the expected value of a log-normal distribution is $\exp(\mu + \sigma^2/2)$. Therefore the imputed values were back-transformed to be on the original scale with a bias correction (David et al., 1986, e.g.). Other methods of transformation could be investigated however there is evidence (e.g. Hippel, 2013) that bias correction often leads to poorer imputed values than using raw data. The log transformation was applied to both the linear and linear mixed model imputed values, and these results were compared with and without transformation. Finally the bias correction was implemented as described above and a further set of imputed values derived for comparison.

In Section 2.5 various criteria for evaluation of imputed values were reviewed. This included calculating the RRMSE and relative bias of the imputed values. In addition to these criteria the impact of the imputation

method will be assessed for summary statistics such as the estimated population mean and variance.

An additional criteria will be consideration of the household structure. Correlation of a continuous variable within households is measured by the ICC, defined as the proportion of total variation due to clustering within households, $ICC = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ (West, Welch, and Galecki, 2007, p.98). This parameter is sometimes referred to as the adjusted ICC, because fixed effects $\boldsymbol{x}$ 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. The ICC can be calculated to assess the impact of an imputation method on within-household clustering:

$$\text{ICC}_{av} = \frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\sigma}_{u,k}^2}{\hat{\sigma}_{u,k}^2 + \hat{\sigma}_{e,k}^2} \tag{3.3}$$

Two final evaluation criteria look at the relative bias for estimating the mean (or similarly for a proportion) and variance. The relative bias for the estimated population mean is:

$$\text{RBias}_{av}(\hat{\bar{Y}}) = \frac{1}{K}\frac{\sum_{k=1}^{K}(\hat{\bar{y}}_k - \bar{y})}{\sum_{k=1}^{K}\bar{y}} \tag{3.4}$$

where $\hat{\bar{y}}_k$ is the sample mean in replicate k and $\bar{y}$ is the sample mean in the full sample with no nonresponse. The relative bias for the estimated

population variance is:

$$\text{RBias}_{av}(\text{var}(y)) = \frac{1}{K}\frac{\sum_{k=1}^{K}(\hat{\text{var}}(y_k) - \text{var}(y))}{\sum_{k=1}^{K}\text{var}(y)} \tag{3.5}$$

where $\hat{\text{var}}(y_k)$ is the sample variance in replicate k and $\text{var}(y)$ is the sample variance in the full sample with no nonresponse.

Another variable relevant for economic policy is the proportion of adults on or below the Federal Minimum Wage (FMW). The impact of the imputation strategy on this variable is investigated. The national minimum wage for 2013/14 is $16.37 per hour (Ombudsman, 2013), with this figure revised each year by a specialist Minimum Wage Panel of the Fair Work Commission. Employees aged under 21 may receive less than the FMW so households with one or more people aged under 21 will be excluded in calculating this quantity. The number of adults in the workforce on minimum wage and their characteristics are used to form the context for debate on the impact of policy changes such as setting the FMW, using data from surveys such as HILDA and the ABS Survey of Income and Housing (Healy and Richardson, 2006). The susceptibility of this measure to the imputation method will be assessed to understand the impact of the imputation method on income in the lower tail of the distribution, rather than just the average. Other points of the distribution function, or quantiles, could also be considered. This thesis only evaluates the proportion at or below the FMW as this is of particular

substantive importance.

In 2004 (aligned with wave 4 of the HILDA survey as used in the simulation study) the federal minimum wage for full-time adult employees was set by the Australian Industrial Relations Commission at $467.40 per week, or $12.30 per hour (Lee and Suardi, 2010). The proportion of adults earning under $12.30 per hour was calculated on the full dataset, and then re-calculated using the imputed hourly wage rate under each imputation method.

3.4 Results

3.4.1 Single-level and multilevel imputation compared to respondent mean

Predictive accuracy: RRMSE of imputed values

Table 3.1 shows the predictive accuracy as measured by the RRMSE of the imputed values, calculated as in Equation (2.16) in Section 2.5 for each imputation method. The accuracy of the imputed values of hourly wage rate has been assessed by comparing the imputed values of non-respondents to the known, true values.

The RRMSE for a respondent mean impute ranges from 56.9% to 67.0%, with larger errors when nonresponse is generated under the MNAR mecha-

Table 3.1: Predictive accuracy (RRMSE %) for imputing *hourly wage rate*

NR model	ρ(%)	Deterministic BLUPs			
		Respmean	SL	SL+	ML
hh MCAR pers MCAR	20	57.0	55.7	55.0	54.9
	50	56.9	55.1	48.8	49.7
	85	58.2	56.7	31.0	36.2
hh MCAR pers MAR	20	57.8	55.9	56.3	55.3
	50	57.9	55.0	50.4	51.2
	85	62.8	61.0	33.7	39.7
hh MCAR pers MNAR	20	65.6	64.4	63.6	63.5
	50	65.1	63.4	55.8	56.6
	85	65.8	64.4	34.3	39.0
hh MNAR pers MCAR	20	61.5	60.2	59.5	59.4
	50	62.5	60.7	53.3	54.1
	85	64.8	63.3	33.6	38.2
hh MNAR pers MNAR	20	67.0	65.8	65.2	64.9
	50	64.3	62.3	55.0	55.4
	85	65.1	63.4	33.9	37.6

Maximum simulation standard error = 0.35

nisms. The deterministic SL BLUP has a small but statistically significant improvement in predictive accuracy compared to respondent mean imputation, around 2-3% for MCAR and MNAR scenarios. The improvement in RRMSE for the SL BLUP over the respondent mean is slightly greater at 3-5% when nonresponse is MAR. This is as expected since the imputation model uses covariates of age group and sex, reflecting the MAR mechanism where males and those under 30 had higher nonresponse. The SL+ and ML BLUP imputed values resulted in no additional improvement over the SL BLUP when $\rho = 20\%$. When $\rho = 50\%$ there is a small improvement in

RRMSE of around 10% for both methods compared to the SL BLUP, but for the highest ρ the improvement in predictive error was much larger. RRMSE decreased by 35-41% across the nonresponse scenarios and by slightly more for SL+ than ML BLUP imputed values. This demonstrates a major improvement in the predictive accuracy of the imputed values for this nonresponse scenario due to the use of household information in the imputation model.

Predictive accuracy: relative bias of imputed values

Table 3.2: Predictive accuracy (relative bias %) for imputing *hourly wage rate*

NR model	ρ(%)	Deterministic BLUPs			
		Respmean	SL	SL+	ML
hh MCAR pers MCAR	20	0.2	0.3	0.2	0.3
	50	0.3	0.2	0.2	0.1
	85	0.2	0.1	0.0	0.0
hh MCAR pers MAR	20	3.1	0.0	0.9	-0.1
	50	3.6	0.7	-0.8	-0.2
	85	2.0	0.1	-0.7	-1.4
hh MCAR pers MNAR	20	-7.5	-7.0	-7.7	-6.5
	50	-7.5	-7.1	-6.4	-5.3
	85	-7.5	-7.2	-2.6	-2.6
hh MNAR pers MCAR	20	-4.3	-4.0	-5.0	-3.2
	50	-5.5	-5.2	-4.4	-3.1
	85	-6.9	-6.6	-1.8	-1.7
hh MNAR pers MNAR	20	-11.1	-10.4	-11.9	-9.2
	50	-11.3	-10.6	-9.5	-7.2
	85	-12.1	-11.4	-3.4	-3.1

Maximum simulation standard error = 0.38

Another measure of predictive accuracy is shown in Table 3.2, the relative

bias of the imputed values. The imputed values of the non-respondents have been compared to the true values of hourly wage rate to determine whether the imputation method tends to over-estimate (positive bias) or under-estimate (negative bias) the true values. Relative biases close to zero are ideal. Under MCAR all of the imputation models have very good bias properties. When the nonresponse was generated under a MAR model, the resulting bias can be seen in the respondent mean impute where imputed values are overestimated by between 2 and 3.6%. The SL, SL+ and ML BLUPs all have small bias (absolute value of 1.4% or less) under MAR.

The further three scenarios where nonresponse is MNAR have high levels of negative bias, that is, there is under-estimation of the missing values by the imputed values. This follows intuitively from the MNAR mechanism as higher wage rates are associated with a higher probability of nonresponse. The deterministic SL BLUPs do not lead to much improvement over the respondent mean. The relative bias of the imputed values does improve under the ML BLUP, with 7-20% bias reduction when $\rho = 20\%$, 25-40% when $\rho = 50\%$, and 64-74% when $\rho = 85\%$. The improvements in bias achieved by using the ML BLUP when nonresponse is MNAR were matched by SL+ when $\rho = 85\%$ but were poorer than the SL BLUP when $\rho = 20\%$, with mixed results when $\rho = 50\%$. These bias reductions are intuitively sensible, as when the nonresponse is dependent on the value of y, it is clear that the

other household member would provide good predictive power, increasing as the ICC increases.

Estimation accuracy - relative bias of estimated mean

The relative bias for the estimated mean of hourly wage rate was calculated as described in Equation (3.4) in Section 3.3.3. A positive relative bias implies that over repeated simulations the mean is over-estimated under this imputation method, and negative biases imply the imputation strategy tends to under-estimate the mean hourly wage rate. The results for the relative bias of the estimated mean reflect the findings relating to the relative bias of the imputed values and are therefore omitted here. The results are included for completeness in Table 7.1 of Appendix A.

Estimation accuracy - intra-cluster correlation

Table 3.3 shows the expected estimated intra-class correlation under each imputation method. This criteria assesses how well the clustering of hourly wage rate within households is retained. When the expected ICC is higher than the true ρ this can be interpreted as the imputed values resulting in hourly wage rates that are more similar within households than in the complete data. Conversely, under-estimates of ρ arise when households have people with hourly wage rates that differ more within households than the complete data after imputation. Respondent mean imputation consistently underestimates ρ under all nonresponse scenarios and ICC levels, varying

Table 3.3: Estimation accuracy for imputing *hourly wage rate* - estimated intra-class correlation

NR model	ρ(%)	Deterministic BLUPs			
		Respmean	SL	SL+	ML
hh MCAR pers MCAR	20	12.7	13.5	24.7	24.2
	50	33.2	35.6	60.7	58.0
	85	56.1	58.7	91.1	87.9
hh MCAR pers MAR	20	14.3	16.4	25.2	25.0
	50	36.4	39.6	64.1	62.1
	85	56.5	59.8	91.1	87.9
hh MCAR pers MNAR	20	11.4	12.4	26.5	26.2
	50	25.9	29.1	61.9	58.7
	85	39.6	43.5	92.8	89.0
hh MNAR pers MCAR	20	10.5	11.6	26.8	26.5
	50	24.6	27.9	62.2	59.1
	85	39.0	43.0	92.9	89.2
hh MNAR pers MNAR	20	9.7	10.8	28.5	28.8
	50	19.6	23.6	63.0	60.0
	85	29.4	34.3	94.2	90.6

Maximum simulation standard error = 0.43

between 25% and close to 65% underestimation of the true ICC. The deterministic SL BLUP is a small improvement but also underestimates the ICC significantly, by between 15% and 60%. The ML BLUP results in overestimates of the ICC, but the values are much more accurate, with the ICC overestimated by between 3% and 50% across the nonresponse mechanisms and ICC levels. The reproduction of the true ICC by the ML approach improves as ρ increases, and is particularly good when $\rho = 85\%$ where the ICC is within 7% of the true value under all nonresponse scenarios. Similar results are achieved by the SL+ impute, though the improvements are not quite as

good, particularly when ρ is high where this method over-estimates the level of clustering within households by more than the ML BLUP does.

Estimation accuracy - relative bias of estimated variance

Table 3.4: Estimation accuracy for imputing *hourly wage rate* - relative bias (%) of estimated variance

NR model		ρ(%)	Deterministic BLUPs			
			Respmean	SL	SL+	ML
hh MCAR pers MCAR		20	-25.0	-23.4	-22.5	-22.6
		50	-25.0	-23.0	-17.7	-17.6
		85	-25.1	-23.4	-6.9	-7.0
hh MCAR pers MAR		20	-23.8	-21.6	-20.1	-20.7
		50	-24.5	-21.9	-15.4	-15.2
		85	-26.8	-24.8	-8.9	-8.0
hh MCAR pers MNAR		20	-37.0	-35.6	-34.6	-34.6
		50	-36.7	-34.9	-29.4	-29.4
		85	-35.9	-34.3	-12.9	-14.4
hh MNAR pers MCAR		20	-31.0	-29.5	-28.4	-28.3
		50	-32.8	-30.9	-24.8	-24.8
		85	-34.4	-32.8	-10.6	-12.1
hh MNAR pers MNAR		20	-40.6	-39.2	-38.2	-38.0
		50	-37.9	-36.2	-30.0	-30.0
		85	-37.7	-36.2	-10.9	-12.8

Maximum simulation standard error = 0.52

The relative bias of the estimated variance (empirical, complete sample variance) for hourly wage was assessed after each imputation method and under the different nonresponse mechanisms, as shown in Table 3.4. As expected the respondent mean imputed values are underestimates of the variance, by 20-40% across the nonresponse and ICC scenarios. The variance is underestimated by the largest magnitude under the MNAR scenarios. That

is, the distribution of the true values is not being maintained by the imputation method.

While still underestimating variance, both of the deterministic BLUP imputed values perform better than using a respondent mean, for all scenarios. Of the three BLUPs, both imputed values incorporating household information are superior to the SL BLUP, however the improvement depends on the level of ρ. For low ρ there is no significant improvement over the SL BLUP, however for moderate ρ the SL+ and ML BLUP estimate variance with between 15-30% less bias than the SL BLUP imputed values across the NR scenarios. When ρ is highest both household-based imputed values are vastly better than the SL BLUP, with around 60-70% reduction in bias compared to the SL BLUP. While the variance is still underestimated when $\rho = 85\%$, it is now by only 7-15% compared to around 25% for the SL BLUP.

A stochastic BLUP would introduce an additional source of variation which may result in improved variance estimates. This will be considered in Chapter 4.

3.4.2 Imputation using log transform

Imputed values were also calculated by using linear models and linear mixed models for the log of hourly wage rate, both with and without bias correction,

as described in Section 3.3.1.

Table 3.5 shows the predictive accuracy as measured by the RRMSE of the respondent mean compared to the SL and ML BLUP imputed values, after log transform and back-transformation, both with and without bias correction (notated 'BC').

Table 3.5: Predictive accuracy (RRMSE %) for imputing *hourly wage rate* using log transform

		BLUP		BLUP log		BLUP log BC	
NR model	ρ(%)	SL	ML	SL	ML	SL	ML
hh MCAR pers MCAR	20	55.7	54.9	56.4	56.2	55.6	54.9
	50	55.1	49.7	55.8	50.6	55.1	50.0
	85	56.7	36.2	57.4	37.6	56.6	39.6
hh MCAR pers MAR	20	55.9	55.3	56.4	56.4	55.7	55.4
	50	55.0	51.2	55.6	51.2	55.0	51.2
	85	61.0	39.7	61.7	42.2	60.9	43.7
hh MCAR pers MNAR	20	64.4	63.5	65.8	65.7	64.4	63.1
	50	63.4	56.6	64.9	59.1	63.5	56.5
	85	64.4	39.0	65.9	42.9	64.5	42.2
hh MNAR pers MCAR	20	60.2	59.4	61.4	61.2	60.2	59.0
	50	60.7	54.1	62.1	56.1	60.8	54.0
	85	63.3	38.2	64.8	41.7	63.4	41.7
hh MNAR pers MNAR	20	65.8	64.9	67.5	67.5	65.9	64.4
	50	62.3	55.4	64.2	58.1	62.4	55.0
	85	63.4	37.6	65.5	41.4	63.6	40.6

The log transform on RRMSE results in slightly poorer RRMSE when comparing the single-level imputed values with and without log transform, and the multilevel imputed values with and without log transform. The bias correction resulted in improved RRMSE, but not as low as the untransformed

SL or ML BLUPs. Therefore there is no gain in the log transform in terms of RRMSE, and without bias correction it leads to poorer imputed values than the untransformed data.

The potential gains of the log transform and subsequent bias correction are also assessed using relative bias as a measure of predictive accuracy in Table 3.6.

Table 3.6: Predictive accuracy (relative bias %) for imputing *hourly wage rate* using log transform

		BLUP		BLUP log		BLUP log BC	
NR model	ρ(%)	SL	ML	SL	ML	SL	ML
hh MCAR persMCAR	20	0·3	0·3	−9·1	−13·1	−0·3	4·9
	50	0·2	0·1	−9·1	−11·2	−0·4	7·2
	85	0·1	0·0	−9·5	−7·5	−0·5	12·2
hh MCAR persMAR	20	0·0	−0·1	−8·9	−13·0	0·0	6·2
	50	0·7	−0·2	−8·3	−10·7	0·6	7·8
	85	0·1	−1·4	−10·0	−8·5	−1·2	10·3
hh MCAR persMNAR	20	−7·0	−6·5	−15·2	−18·7	−7·4	−2·8
	50	−7·1	−5·3	−15·2	−16·0	−7·4	0·2
	85	−7·2	−2·6	−15·7	−10·1	−7·7	7·4
hh MNAR persMCAR	20	−4·0	−3·2	−12·7	−16·1	−4·4	0·5
	50	−5·2	−3·1	−13·7	−14·2	−5·6	2·5
	85	−6·6	−1·7	−15·2	−9·3	−7·1	8·4
hh MNAR persMNAR	20	−10·4	−9·2	−18·1	−21·1	−10·7	−6·4
	50	−10·6	−7·2	−18·3	−17·9	−10·9	−2·7
	85	−11·4	−3·1	−19·4	−10·6	−11·9	5·7

The log transformation introduces very large negative biases, more so than what was present in the untransformed data, both for SL and ML BLUPs. The bias correction works well at undoing this for the SL BLUP,

with the results after transformation and bias correction closely matching the bias levels for the untransformed data. The ML BLUP with a log transform achieves mixed results, the bias correction generally over-estimating the imputed values. For the MCAR and MAR scenarios, the imputed values are too high, with much larger relative bias than the untransformed ML BLUP. Under MNAR the correction works well, but still resulting in mixed results, particulary under MNAR scenarios with high ρ where the bias corrected data over-shoots the true values.

The relative bias of the estimated mean after imputation was calculated with the same imputation methods as described above. The results of this analysis of the estimated mean after imputation reflected the problems above. There was further negative bias when using log transformed compared to raw data imputed values, which was corrected with the bias factor for the SL model. The same mixed results were achieved when using the ML BLUP with the bias correction. That is poorer bias under MCAR or MAR, or MNAR with high ρ, and a bias improvement for the MNAR scenarios with low or moderate ρ. The results are in Table 7.2, in Appendix B.

Table 3.7 shows the estimated intra-class correlation of hourly wage rate, averaged over $K = 250$ replicates, for imputed values using the SL BLUP and ML BLUP with and without the log transformation and with a bias correction.

Table 3.7: Estimation accuracy for imputing *hourly wage rate* - estimated intra-class correlation with log transform

		BLUP		BLUP log		BLUP log BC	
NR model	ρ(%)	SL	ML	SL	ML	SL	ML
hh MCAR persMCAR	20	13·5	24·2	13·1	21·0	13·5	23·2
	50	35·6	58·0	35·0	53·4	35·5	56·9
	85	58·7	87·9	57·9	86·2	58·6	87·3
hh MCAR persMAR	20	16·4	25·0	17·0	23·9	16·3	23·8
	50	39·6	62·1	39·1	57·3	39·5	60·4
	85	59·8	87·9	59·2	86·5	59·8	87·5
hh MCAR persMNAR	20	12·4	26·2	11·4	20·4	12·3	24·0
	50	29·1	58·7	27·4	51·2	28·9	56·9
	85	43·5	89·0	41·0	85·4	43·1	88·4
hh MNAR persMCAR	20	11·6	26·5	10·3	19·5	11·5	23·7
	50	27·9	59·1	26·0	50·9	27·7	56·9
	85	43·0	89·2	40·4	85·7	42·5	88·7
hh MNAR persMNAR	20	10·8	28·8	8·9	18·6	10·7	23·9
	50	23·6	60·0	20·7	49·0	23·3	56·7
	85	34·3	90·6	30·5	86·0	33·6	90·1

The log transformation has only a small impact on the estimated ICC when using a SL BLUP impute. The estimated ICC was slightly worse after the log transform than on raw data, particularly across the three MNAR nonresponse models when $\rho = 85\%$. However, the log transform results in excellent reproduction of intra-household clustering when using the ML BLUP across all scenarios. The bias correction improves the SL BLUP imputed value's estimation of ICC back to in line with the raw data, but it is a backwards step for the ML BLUP where the estimated ICCs were too high.

Finally the variance is evaluated in Table 3.8, with the relative bias of the

Table 3.8: Estimation accuracy for imputing *hourly wage rate* - relative bias (%) of estimated variance with log transform

				BLUP log		BLUP log BC	
NR model	ρ(%)	SL	ML	SL	ML	SL	ML
hh MCAR pers MCAR	20	−23·4	−22·6	−23·4	−22·3	−23·7	−22·5
	50	−23·0	−17·6	−23·1	−19·1	−23·4	−17·4
	85	−23·4	−7·0	−23·5	−10·6	−23·8	−3·8
hh MCAR pers MAR	20	−21·6	−20·7	−21·4	−20·2	−21·8	−20·6
	50	−21·9	−15·2	−21·9	−17·1	−22·3	−15·5
	85	−24·8	−8·0	−24·7	−11·0	−25·1	−4·7
hh MCAR pers MNAR	20	−35·6	−34·6	−35·6	−34·6	−35·8	−34·7
	50	−34·9	−29·4	−35·1	−31·3	−35·2	−29·4
	85	−34·3	−14·4	−34·4	−19·1	−34·6	−10·9
hh MNAR pers MCAR	20	−29·5	−28·3	−29·5	−28·5	−29·7	−28·5
	50	−30·9	−24·8	−31·1	−27·0	−31·2	−25·0
	85	−32·8	−12·1	−33·0	−16·8	−33·2	−8·3
hh MNAR pers MNAR	20	−39·2	−38·0	−39·3	−38·5	−39·5	−38·4
	50	−36·2	−30·0	−36·3	−32·5	−36·5	−30·4
	85	−36·2	−12·8	−36·4	−18·4	−36·6	−9·0

estimated variance shown for the raw data, after a log transform, and with the log transform and a subsequent bias correction. For SL BLUP imputed values there is not much to separate the methods in terms of variance, with neither the transformation or bias correction having much of an affect. For the ML BLUP imputed values this time the log transformation with bias correction is an improvement on the untransformed data. The variance is under-estimated by a lesser amount with the introduction of the log transform with correction when $\rho = 85\%$ and otherwise similar to the original results for the untransformed data.

3.4.3 Proportion below Federal Minimum Wage

Table 3.9: Various estimates of the percentage of adults on or below FMW

NR model		ρ(%)	Actual	Respmean	BLUP SL	BLUP ML	BLUP log SL	BLUP log ML
hh MCAR pers MCAR		20	7.7	5.8	5.8	5.8	5.8	5.8
		50	6.8	5.1	5.1	5.7	5.1	5.4
		85	6.7	5.0	5.0	7.2	5.0	6.0
hh MCAR pers MAR		20	7.7	5.8	5.8	5.8	5.8	5.8
		50	6.8	5.0	5.0	6.0	5.0	5.3
		85	6.7	4.8	4.8	7.5	4.8	6.2
hh MCAR pers MNAR		20	7.7	6.0	6.0	6.0	6.0	6.0
		50	6.8	5.3	5.3	5.7	5.3	5.5
		85	6.7	5.1	5.1	7.0	5.1	6.1
hh MNAR pers MCAR		20	7.7	5.9	5.9	5.9	5.9	5.9
		50	6.8	5.2	5.2	5.7	5.2	5.5
		85	6.7	5.1	5.1	7.0	5.1	6.1
hh MNAR pers MNAR		20	7.7	6.1	6.1	6.1	6.1	6.1
		50	6.8	5.4	5.4	5.8	5.4	5.6
		85	6.7	5.3	5.3	6.9	5.3	6.1

Table 3.9 shows the expected values of the estimated proportion of people on or below the FMW, using the various imputation methods. There are no differences between the estimates based on the respondent mean impute and SL BLUP imputed values, under any scenario. Both imputation methods resulted in between 25% and 40% underestimation of the percentage of people on or below the FMW, implying that these nonrespondents had their hourly wage rate over-inflated by the imputation method. The ML BLUP gave no improvement over either method under low ICC, but under moderate

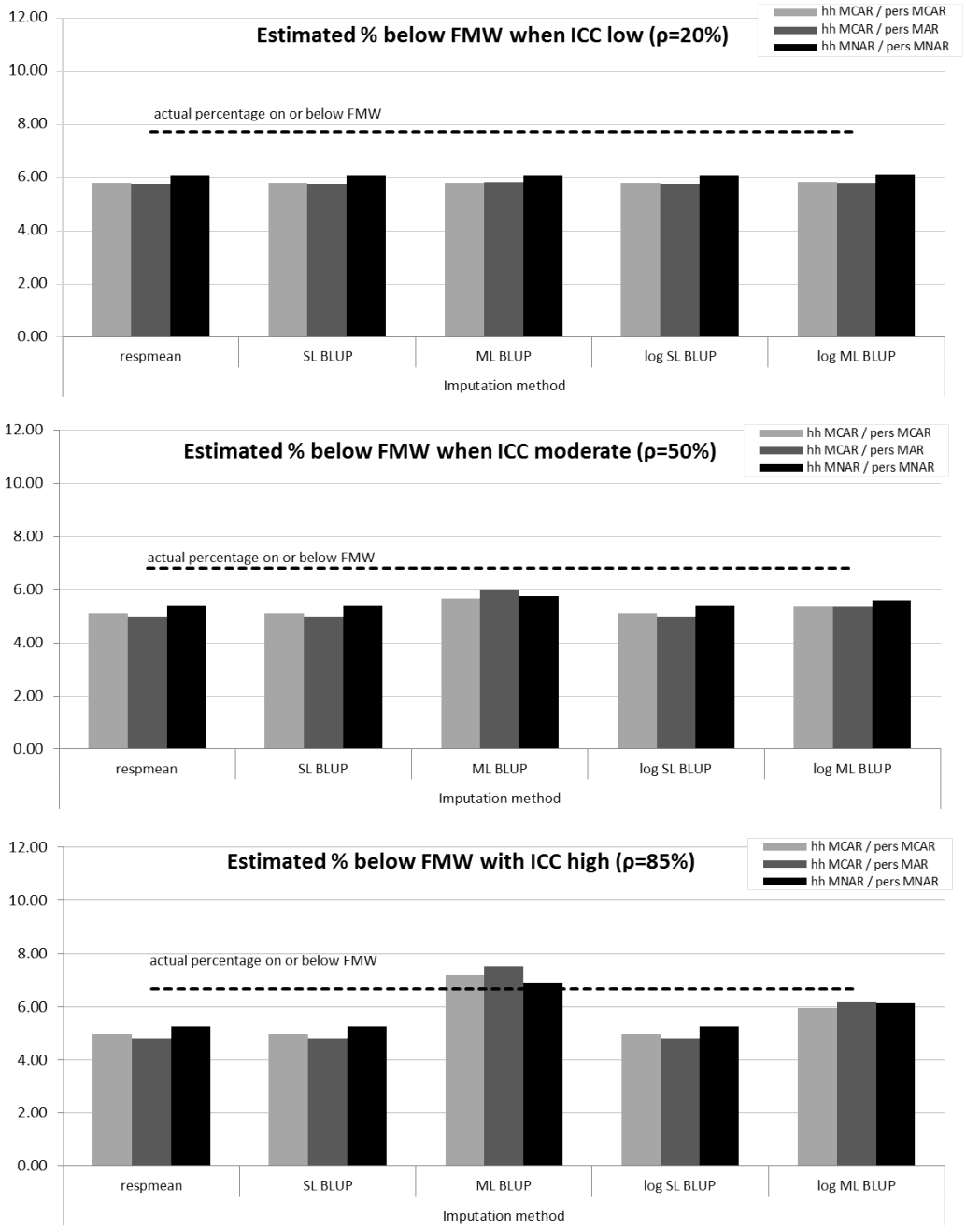

Figure 3.1: Simulation study - estimated percentage on or below Federal Minimum wage

ICC the ML BLUP resulted in substantial reductions in the relative bias of the estimate. The percentage of people on or under the FMW was still underestimated, but by a smaller amount. For example when household nonresponse was MCAR and persons MAR the true percentage of people on or below the FMW was 7.0% , which, under respondent mean imputation and SL BLUP imputation, was under-estimated at 5.0%, but under the ML BLUP was estimated to be 5.9%. The corresponding relative bias was reduced from -36.9% to -15.5%.

The findings are also displayed in Figure 3.1, which shows the estimated percentage of people below FMW when $\rho = 20\%$, $\rho = 50\%$ and $\rho = 85\%$ respectively. The percentage of people on or below FMW is underestimated using all imputation methods, with the exception of the use of ML method when $\rho = 85\%$.

3.5 Summary of Chapter 3

The main question posed in this chapter was whether imputations using information about other people within a household do better than the more standard use of a single-level model with only information about the non-respondent themselves. The answer is yes, particularly when nonresponse is informative both of households and within households. The improvement

over the SL BLUP increases as ρ increases. Improvements in imputed values were achieved whether the information was incorporated in a single model via an additional covariate, or by the use of a two-level model, although overall the ML BLUP achieved slightly better results.

The ML BLUP and SL BLUP with household respondent improved predictive accuracy as measured by RRMSE, with the improvement depending on the ICC. Both imputation methods incorporating household information resulted in a significant improvement in predictive accuracy compared to the SL BLUP. While there was no reduction in RRMSE for low ρ, the SL BLUP with household respondent and ML BLUP resulted in a reasonable improvement in predictive accuracy for moderate ρ, and a large improvement for high ρ, consistent in magnitude across each of the nonresponse scenarios. The single-level model incorporating co-householder's income had slightly lower RRMSE than the two-level model. In relation to bias the household imputation methods had no impact under MCAR or MAR where the bias is already very low, but the ML BLUP reduced the bias of the imputed values under each MNAR mechanism, underestimating income by a smaller amount than the SL BLUP. The improvement seen in the ML BLUP again depended on the ICC, with the largest improvements occurring with higher ρ. The single-level model incorporating co-householder's income was poorer than the two-level model for relative bias, particularly under low and moderate

levels of clustering.

The ML BLUP was the standout imputation method for reproducing ICC. The SL BLUP under-estimated ICC significantly for all values of ρ. While both household imputation methods induced too much clustering, the ML BLUP did so by less, and achieved ICCs closer to the true values, particularly for high ρ. All three BLUPs under-estimated the variance by a similar amount when $\rho = 20\%$. For moderate ρ the ML BLUP and SL BLUP with covariate estimated variance with less bias than the SL BLUP imputed values. When ρ is highest both household imputed values reduce bias in variance estimation compared to the SL BLUP.

The simulation study included BLUPs based on linear models for $log(Y)$, both with and without bias correction. The log transformation resulted in bias issues for both SL and ML BLUP imputed values. These were mostly addressed by the bias correction for the SL BLUP impute, but some bias issues remained for the ML BLUP imputed values after back transformation and bias correction. This was also seen in the slightly worse RRMSE for the ML BLUP imputed values with the log transformation after bias correction, with poorer accuracy when $\rho = 85\%$.

Estimation of ICC was compared across the imputation methods after log transformation compared to raw data. The SL BLUP with the log transform including a bias correction was no improvement on the SL BLUP imputed val-

ues with untransformed data. ML BLUP imputed values after log transform were an improvement for estimation of ICC on the untransformed results. The bias correction made the clustering further from the true values, but was still an improvement on untransformed data, and much better than the SL BLUP imputed values.

The log transformation had no impact on variance estimation for the SL BLUP imputed values, however, for the ML BLUP imputed values, variance estimation was slightly worse using the log transformed data, but slightly better than untransformed data when the bias correction was applied.

In summary, based on the evaluation criteria considered above, the log transformation is worth considering but should be used with caution. Imputed values based on log transformed data should not be used without bias correction, and for the ML BLUP further investigation is needed to determine whether an improved bias correction is possible for multilevel data with high levels of clustering.

This chapter included an evaluation of respondent mean, SL BLUP and ML BLUP imputed values on a range of criteria relevant to household surveys. The respondent mean and SL BLUP imputed values were shown to substantially under-estimate clustering within households. The ML BLUP and SL BLUP with co-householder are much better, although they tend to impute too similar values within households, with the ML BLUP being the

slightly better method. All imputation methods under-estimated variability, however stochastic BLUP imputed values would be expected to better reproduce variability, and may also more accurately reproduce clustering. The use of stochastic and multiple imputed values will be the topic of the following chapter.

Chapter 4

Imputation of Continuous Data using Stochastic Linear Models and Linear Mixed Models

4.1 Introduction

Single imputation methods based on deterministic linear and linear mixed models were found to do poorly in Chapter 3 in reproducing variation and level of clustering within households. Intuitively a sample using a single deterministic impute based on a linear impute will contain less dispersion than would have been seen had the survey participants provided responses. Each of the deterministic linear imputation methods considered in Chapter 3 (re-

spondent mean, both SL BLUPs and ML BLUP) underestimated population variance for hourly wage. The bias of this variance was between 20% and 30% when using a respondent mean or SL BLUP impute, and between 5% and 27% for ML BLUP, with the better variance estimates achieved when ICC was highest. Both the respondent mean and SL BLUP impute also resulted in underestimation of the ICC, however the household imputation methods, the SL BLUP with co-householder income, and ML BLUP were able to counter this and instead slightly over-corrected the within-household correlation.

The findings of Chapter 3 with regard to variance and clustering point to the use of stochastic rather than deterministic imputation methods. Stochastic imputed values have a random mechanism potentially resulting in a different impute each time the imputation method is applied. This requires specification of a distributional model for the variable of interest for use in imputation. Random draws can then be generated from a fitted model based on the available data.

Stochastic imputed values extend naturally to repeated imputation. The mechanism used to derive a single impute may be repeated several times to create a number of imputed values for each single missing data item. These multiple imputed values, even if not proper in the Bayesian sense, can be used to derive improved estimators of the variance of population quantities

of parameters.

This chapter will explore the impact of using household imputation methods with a stochastic component, and whether this solves the issues relating to clustering and variance of deterministic imputed values described in Chapter 3. Section 4.2 will include a description of the theory for deriving one or more stochastic imputed values under a single and multilevel linear model. Multiple imputed values will be calculated and combined as described in Section 2.4. The stochastic imputation methods are assessed in Section 4.3 using an extension of the simulation study in Chapter 3, imputing *hourly wage rate* in the HILDA survey. This section examines the relative performance of the different stochastic methods, as well as their advantages over deterministic methods. The evaluation will include their relative performance for calculating the percentage of people on or under the Federal Minimum Wage (FMW).

4.2 Conditional Distribution of Missing Values

A stochastic impute can be calculated by assuming the vectors of the variable of interest for all household respondents are independent multivariate nor-

mally distributed random variables. Imputed values can then be drawn from the conditional distribution of the missing values given the observed data, substituting estimators for unknown parameters. We would like to draw from the conditional distribution of the missing data Y_{ij} given the observed Y_o. Assume the missing $Y_{i'j} \sim$ N and the vector of respondents $\mathbf{Y_o} \sim$ MVN (multivariate normal).

Given two variables A and $\mathbf{B}$ where

$$\begin{pmatrix} A \\ \boldsymbol{B} \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} \mu_A \\ \boldsymbol{\mu_B}, \end{pmatrix} \begin{bmatrix} \sigma_A^2 & \mathbf{\Sigma_{AB}} \\ \mathbf{\Sigma_{AB}^T} & \mathbf{\Sigma_{BB}} \end{bmatrix} \right)$$

it is well known that the conditional distribution of $A|\mathbf{B} = \mathbf{b}$ is also MVN with:

$$\text{E}(A|\mathbf{B} = \mathbf{b}) = \mu_A + \mathbf{\Sigma_{AB}\Sigma_{BB}^{-1}}(\mathbf{b} - \mu_\mathbf{B})$$

$$\text{Cov}(A|\mathbf{B} = \mathbf{b}) = \sigma_A^2 - \mathbf{\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{AB}^T}$$

Under Model (3.1), the expectation of the missing data conditional on the observed is therefore

$$\text{E}(Y_{i'j}|\mathbf{Y_o} = \mathbf{y_o}) = E(Y_{i'j}) + \text{Cov}(Y_{i'j}, \mathbf{y_o})V^{-1}(\mathbf{y_o})(\mathbf{y_o} - E(\mathbf{y_o}))$$

Now $\text{Cov}(Y_{ij}, Y_{i'j}) = \rho$, and $\text{Cov}(Y_{i'j}, Y_{ij'}) = 0 \quad \text{for} j \neq j'$. Therefore

$$\text{E}(Y_{i'j}|\mathbf{Y_o} = \mathbf{y_o}) = E(Y_{i'j}) + \text{Cov}(Y_{i'j}, \mathbf{y_{o,j}})V^{-1}(\mathbf{y_{o,j}})(\mathbf{y_{o,j}} - E(\mathbf{y_{o,j}}))$$

The variance matrix $V(\mathbf{y_{o,j}})$ is a square matrix of size n_{rj} as follows:

$$V(\mathbf{y_{o,j}}) = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}$$

The inverse of a matrix in this form (e.g. Healy, 2000, p40-41) of size n is

$$\begin{pmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{pmatrix}$$

where

$$a = \frac{1 + (n-2)\rho}{(1-\rho)(1+(n-1)\rho)} \qquad \text{and}$$

$$b = \frac{-\rho}{(1-\rho)(1+(n-1)\rho)}$$

The inverse can also be written in matrix form as $(a-b)\mathbf{I} + b\mathbf{J}$ where I is the identity matrix of size n and $\mathbf{J}$ is a square matrix of 1's. Let $\mathbf{1}$ be a vector of 1's of length n_{rj}, $\mathbf{I}$ the identity matrix of size n_{rj} and $\mathbf{J}$ be a n_{rj} square

matrix of 1's. The expectation then becomes:

$$\begin{aligned}
\mathrm{E}\left(Y_{i'j}|\mathbf{Y_o}=\mathbf{y_o}\right) &= \mathbf{x}_{i'j}^{T}\boldsymbol{\beta}+\sigma^{2}\rho\mathbf{1^T}\sigma^{-2}\left(\frac{1}{1-\rho}\mathbf{I}-\frac{\rho}{\left(1-\rho\right)\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{J}\right)\\
&\quad\times\left(\mathbf{y_{o,j}}-\mathbf{x_{o,j}}\boldsymbol{\beta}\right)\\
&= \mathbf{x}_{i'j}^{T}\boldsymbol{\beta}+\frac{\rho}{1-\rho}\left(\mathbf{1^T}\mathbf{I}-\frac{\rho}{\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{1J}\right)\left(\mathbf{y_{o,j}}-\mathbf{x_{o,j}}\boldsymbol{\beta}\right)\\
&= \mathbf{x}_{i'j}^{T}\boldsymbol{\beta}+\frac{\rho}{1-\rho}\left(\mathbf{1^T}-\frac{\rho n_{rj}}{\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{1^T}\right)\left(\mathbf{y_{o,j}}-\mathbf{x_{o,j}}\boldsymbol{\beta}\right)
\end{aligned}$$

$$\begin{aligned}
&= \mathbf{x}_{i'j}^{T}\boldsymbol{\beta}+\frac{\rho}{1-\rho}\left(\mathbf{1^T}-\frac{\rho n_{rj}}{\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{1^T}\right)\left(\mathbf{y_{o,j}}-\mathbf{x_{o,j}}\boldsymbol{\beta}\right)\\
&= \mathbf{x}_{i'j}^{T}\boldsymbol{\beta}+\frac{\rho}{1-\rho}\left(\frac{1+\left(n_{rj}-1\right)\rho-\rho n_{rj}}{\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{1^T}\right)\left(\mathbf{y_{o,j}}-\mathbf{x_{o,j}}\boldsymbol{\beta}\right)\\
&= \mathbf{x}_{i'j}^{T}\boldsymbol{\beta}+\rho\left(1+\left(n_{rj}-1\right)\rho\right)^{-1}\mathbf{1^T}\left(\mathbf{y_{o,j}}-\mathbf{x_{o,j}}\boldsymbol{\beta}\right)
\end{aligned}$$

where $\mathbf{x_{o,j}}$ is the matrix of covariates for the respondents in household j. When there is only one respondent in the household $n_{rj}=1$ the expectation

simplifies to $\mathbf{x}_{i'j}^T\boldsymbol{\beta} + \rho\left(y_{i,j} - \mathbf{x}_{i,j}^T\boldsymbol{\beta}\right)$. The variance is given by:

$$\begin{aligned}
\mathrm{Var}\left(Y_{i'j}|\mathbf{Y_o} = \mathbf{y_o}\right) &= \sigma^2 - \mathrm{Cov}\left(Y_{i'j}, \mathbf{y_o}\right)\mathbf{V^{-1}}\left(\mathbf{y_o}\right)\mathrm{Cov}\left(\mathbf{y_o}, Y_{i'j}\right) \\
&= \sigma^2 - \mathrm{Cov}\left(Y_{i'j}, \mathbf{y_j}\right)\mathbf{V^{-1}}\left(\mathbf{y_{o,j}}\right)\mathrm{Cov}\left(\mathbf{y_{o,j}}, Y_{i'j}\right) \\
&= \sigma^2 - \left(\sigma^2\rho\mathbf{1^T}\right)\mathbf{V^{-1}}\left(\mathbf{y_{o,j}}\right)\left(\sigma^2\rho\mathbf{1^T}\right)^T \\
&= \sigma^2\left[1 - \sigma^2\rho^2\mathbf{1^T}\mathbf{V^{-1}}\left(\mathbf{y_{o,j}}\right)\mathbf{1}\right] \\
&= \sigma^2\left[1 - \sigma^2\rho^2\mathbf{1^T}\sigma^{-2}\left(\frac{1}{1-\rho}\mathbf{I} - \frac{\rho}{\left(1-\rho\right)\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{J}\right)\mathbf{1}\right] \\
&= \sigma^2\left[1 - \frac{\rho^2}{1-\rho}\left(\mathbf{1^T1} - \frac{\rho}{\left(1+\left(n_{rj}-1\right)\rho\right)}\mathbf{1^TJ1}\right)\right] \\
&= \sigma^2\left[1 - \frac{\rho^2}{1-\rho}\left(n_{rj} - \frac{\rho n_{rj}^2}{\left(1+\left(n_{rj}-1\right)\rho\right)}\right)\right] \\
&= \sigma^2\left[1 - \frac{n_{rj}\rho^2}{1-\rho}\left(1 - \frac{\rho n_{rj}}{\left(1+\left(n_{rj}-1\right)\rho\right)}\right)\right] \\
&= \sigma^2\left[1 - \frac{n_{rj}\rho^2}{1-\rho}\left(\frac{1+\left(n_{rj}-1\right)\rho - \rho n_{rj}}{\left(1+\left(n_{rj}-1\right)\rho\right)}\right)\right] \\
&= \sigma^2\left[1 - n_{rj}\rho^2\left(1+\left(n_{rj}-1\right)\rho\right)^{-1}\right]
\end{aligned}$$

This can also be generalised to several missing Y and several observed Y within a household. When there is only one respondent in the household $n_{rj} = 1$ and the variance simplifies to $\sigma^2\left(1-\rho^2\right)$.

For example, for two person households with one respondent, $n_{rj} = 1$, the SL BLUP has $\rho = 0$ and stochastic imputed values are drawn as follows:

$$Y_{i'j} \sim N\left(\mathbf{x}_{i'j}\hat{\beta}, \hat{\sigma}^2\right). \tag{4.1}$$

For the ML BLUP a draw is taken from:

$$Y_{i'j} \sim N(\mathbf{x}_{i'j}\hat{\beta} + \hat{\rho}(Y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}), (1 - \hat{\rho}^2)\hat{\sigma}^2) \tag{4.2}$$

For the single-level model $\hat{\beta}$ and $\hat{\sigma}^2$ are the ordinary least squared estimators. The multilevel model uses estimates for $\hat{\boldsymbol{\beta}}$, ρ and $\hat{\sigma}^2$ produced by IGLS using the PROC MIXED procedure in SAS v9.2. Estimation of these parameters implicitly assumes a MAR nonresponse mechanism.

The stochastic nature of these methods is expected to result in the imputed dataset having more realistic variances and intra-class correlations, potentially at the expense of worse predictive accuracy at the individual level. Either single or multiple stochastic imputed values are able to be drawn from these distributions. When multiple imputed values are drawn the imputed values are combined as described in Section 2.4.4.

4.3 Simulation study

The simulation study in Section 3.3 of the previous chapter using hourly wage rate from HILDA (Watson, 2008) was extended by deriving stochastic SL BLUP imputed values using (4.1) and stochastic ML BLUP imputed values using (4.2). Five stochastic imputation methods were added to the simulation study from Chapter 3 for imputing missing Y_{ij}:

- *Stochastic Single-level BLUP:* single draw from the conditional distribution of the missing, given the observed data, as in (4.1) with age by sex as auxiliary variables (notated stochastic SL);

- *Stochastic Single-level BLUP:* single draw from the conditional distribution of the missing, given the observed data, as in (4.1) with age by sex as auxiliary variables and an additional covariate given by the variable of interest for an item respondent from the same household (notated stochastic SL+);

- *Stochastic Multilevel BLUP:* single draw from the conditional distribution of the missing, given the observed data, as in (4.2) with age by sex as auxiliary variables (notated stochastic ML);

- *Multiply Imputed Single-level BLUP:* multiple draws from the conditional distribution of the missing, assuming the observed, as in (4.1) with age by sex as auxiliary variables (notated MI SL);

- *Multiply Imputed Multilevel BLUP:* multiple draws from the conditional distribution of the missing, assuming the observed, as in (4.2) with age by sex as auxiliary variables (notated MI ML).

As in the previous chapter, the BLUP imputation models used age by sex as explanatory variables 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 or item respondent.

As with the deterministic methods, the initial stochastic imputed values were calculated using untransformed data. The log transformation was then applied and the linear and linear mixed stochastic imputed values calculated with and without this transformation.

As in Subsection 3.3.3, ICCs of $\rho = 19.4\%$, $\rho = 50.0\%$ and $\rho = 85.0\%$ were used to assess the imputation methods under low, moderate and high ICC.

In consideration of the required number of imputed values, Rubin (1987b, p114) described the efficiency of an estimate based on m imputed values as approximately $(1 + \frac{\gamma}{m})^{-1}$, where γ is the rate of missing information for the quantity being estimated. The efficiencies for various values of γ and m are shown below. The simulation study has been carried out with $m = 30$ imputed values, which has a high level of efficiency for across varying levels of missing information.

Table 4.1: Efficiency for various rates of missing information and number of multiple imputed values

	γ				
m	0.1	0.25	0.5	0.7	0.9
2	95	89	80	74	69
3	97	92	86	81	77
5	98	95	91	88	85
10	99	98	95	93	92
30	100	99	98	98	97
50	100	100	99	99	98

4.4 Results

This section contains the results of a comparison of the BLUP under single and multilevel linear imputation models using stochastic methods compared to the deterministic methods evaluated in Chapter 3 . Respondent mean imputed values were used as a point of comparison. The simulation study was also carried out with a log transform for hourly wage rate. A log transform of the outcome was performed prior to imputation, and the imputed values back-transformed to be on the original scale with a bias correction as described in Section 3.3.1. This correction was applied to address the bias issues found in the deterministic imputed values, which were in part resolved by the correction.

Predictive Accuracy was assessed at an individual level by calculating the RRMSE of prediction as in (2.16), and the relative bias as in (2.17), averaged over the 250 replicates. Estimation accuracy was assessed for means, vari-

ance and intra-household correlation (ICC) under the different nonresponse models for each imputation method, each averaged over 250 replicates. When multiple stochastic imputed values were calculated, they were combined by averaging over the $m = 30$ imputed values in each of the 250 replicates.

4.4.1 Single stochastic linear and linear mixed imputed values compared to deterministic imputed values

Table 4.2: RRMSE (%) - imputation using single stochastic BLUPs compared to deterministic BLUPs

NR model		ρ(%)	Respmean	Deterministic SL	Deterministic SL+	Deterministic ML	Stochastic SL	Stochastic SL+	Stochastic ML	Stochastic log SL	Stochastic log ML
hh	MCAR	20	57.0	55.7	55.0	54.9	78.9	77.6	77.7	75.7	76.7
pers	MCAR	50	56.9	55.1	48.8	49.7	78.2	69.0	70.3	75.3	70.2
		85	58.2	56.7	31.0	36.2	80.3	43.7	50.7	76.9	54.2
hh	MCAR	20	57.8	55.9	56.2	55.3	80.2	81.8	81.4	76.1	80.2
pers	MAR	50	57.9	55.0	50.4	51.2	79.5	70.8	72.1	75.9	71.0
		85	62.8	61.0	33.7	39.7	83.2	45.0	52.2	79.7	55.6
hh	MCAR	20	65.6	64.4	63.6	63.5	80.5	77.4	77.3	78.7	78.0
pers	MNAR	50	65.1	63.4	55.8	56.6	79.6	67.6	68.6	78.0	70.9
		85	65.8	64.4	34.3	39.0	81.4	41.8	47.7	79.0	53.7
hh	MNAR	20	61.5	60.2	59.5	59.4	79.4	74.9	74.9	76.8	76.4
pers	MCAR	50	62.5	60.7	53.3	54.1	78.7	65.9	67.0	76.6	70.0
		85	64.8	63.3	33.6	38.2	80.8	41.2	47.0	78.3	53.6
hh	MNAR	20	67.0	65.8	65.2	64.9	79.9	74.8	74.6	78.2	76.8
pers	MNAR	50	64.3	62.3	55.0	55.4	77.6	63.8	64.4	75.6	68.1
		85	65.1	63.4	33.9	37.6	79.0	39.1	43.7	76.4	51.4

maximum simulation standard error = 0.35

Table 4.2 shows the RRMSE under the five nonresponse mechanisms de-

scribed in Section 3.3.2 and for three values of ρ. The three deterministic SL and ML BLUPs of Chapter 3 are presented alongside the results for the stochastic SL and ML BLUPs. The results of a log transform on the SL and ML BLUP are also presented.

The stochastic BLUPs result in a higher RRMSE than the deterministic methods a a result of the random component in the method. Both imputation methods which incorporate the household information (SL+ and ML) demonstrate improvements over the SL BLUP. The improvement is small when $\rho = 20\%$, and mostly occurs when nonresponse is informative. The improvement in accuracy increases with ρ, resulting in 9-18% reduction in RRMSE in both the SL+ and ML imputation methods for moderate ρ and 37-45% for the highest ICC level with ML BLUP, and slightly more for SL+ (46-51%). This is consistent with the comparison of deterministic BLUPs. After the log transformation improvement in RRMSE by introducing a ML impute is of a similar magnitude to that seen without the log transform.

The relative bias of the imputed values are shown in Table 3.2. The introduction of a stochastic component in the imputed values had almost no impact on bias. However, there were bias problems when the log transform was employed under MCAR and MAR nonresponse models. The bias in the imputed values is also reflected in the bias of the estimated mean (Table 7.1 of Appendix B). The untransformed BLUPs have much better bias properties

Table 4.3: Relative bias (%) - imputation using single stochastic BLUPs compared to deterministic BLUPs

NR model		ρ(%)	Respmean	Deterministic SL	Deterministic SL+	Deterministic ML	Stochastic SL	Stochastic SL+	Stochastic ML	Stochastic log SL	Stochastic log ML
hh MCAR pers MCAR		20	0.2	0.3	0.2	0.3	0.5	0.4	0.3	9.4	14.6
		50	0.3	0.2	0.2	0.1	0.4	0.3	0.1	9.3	15.0
		85	0.2	0.1	0.0	0.0	0.3	0.2	0.0	9.4	16.2
hh MCAR pers MAR		20	3.1	0.0	0.9	-0.1	-0.1	0.7	0.1	9.7	17.1
		50	3.6	0.7	-0.8	-0.2	0.6	-0.9	0.0	10.2	15.4
		85	2.0	0.1	-0.7	-1.4	0.1	-0.6	-1.2	8.5	13.9
hh MCAR pers MNAR		20	-7.5	-7.0	-7.7	-6.5	-7.1	-7.7	-6.5	1.2	5.5
		50	-7.5	-7.1	-6.4	-5.3	-7.2	-6.5	-5.3	1.1	7.0
		85	-7.5	-7.2	-2.6	-2.6	-7.3	-2.6	-2.6	1.0	10.9
hh MNAR pers MCAR		20	-4.3	-4.0	-5.0	-3.2	-3.8	-4.8	-3.0	4.7	9.3
		50	-5.5	-5.2	-4.4	-3.1	-5.1	-4.3	-3.0	3.3	9.6
		85	-6.9	-6.6	-1.8	-1.7	-6.5	-1.7	-1.6	1.8	12.1
hh MNAR pers MNAR		20	-11.1	-10.4	-11.9	-9.2	-10.3	-11.8	-9.1	-2.5	1.1
		50	-11.3	-10.6	-9.5	-7.2	-10.4	-9.4	-7.1	-2.7	3.7
		85	-12.1	-11.4	-3.4	-3.1	-11.3	-3.4	-3.1	-3.6	8.9

maximum simulation standard error = 0.28

in both the MCAR and MAR scenarios. Under a log transform the mean is over-estimated by both the SL and ML log-transformed BLUP in all but the last MNAR scenario.

Table 4.4 shows the estimated intra-class correlation after imputation. The stochastic SL BLUP is poorer than the deterministic method, under-estimating the ICC even further. However, both stochastic household imputation methods impute the clustering levels remarkably well. Unlike the deterministic methods, they only induce a slight over-clustering, and only under

Table 4.4: Expected value of estimated ICC (%) - imputation using sing stochastic BLUPs compared to deterministic BLUPs

NR model		ρ(%)	Respmean	Deterministic SL	Deterministic SL+	Deterministic ML	Stochastic SL	Stochastic SL+	Stochastic ML	Stochastic log SL	Stochastic log ML
hh	MCAR	20	12.7	13.5	24.7	24.2	10.4	19.2	18.9	10.6	18.7
pers	MCAR	50	33.2	35.6	60.7	58.0	27.4	50.0	47.3	28.3	47.3
		85	56.1	58.7	91.1	87.9	45.0	84.9	79.9	46.7	79.3
hh	MCAR	20	14.3	16.4	25.2	25.0	12.8	19.3	19.2	12.3	17.8
pers	MAR	50	36.4	39.6	64.1	62.1	30.3	52.9	51.0	31.4	50.9
		85	56.5	59.8	91.1	87.9	46.5	85.4	81.1	48.3	81.1
hh	MCAR	20	11.4	12.4	26.5	26.2	9.5	21.2	21.0	10.1	19.5
pers	MNAR	50	25.9	29.1	61.9	58.7	22.1	52.5	49.6	23.1	47.3
		85	39.6	43.5	92.8	89.0	33.1	88.1	83.0	34.9	80.6
hh	MNAR	20	10.5	11.6	26.8	26.5	9.0	21.9	21.6	9.8	19.7
pers	MCAR	50	24.6	27.9	62.2	59.1	21.3	53.3	50.3	22.5	48.0
		85	39.0	43.0	92.9	89.2	32.8	88.4	83.3	34.5	80.9
hh	MNAR	20	9.7	10.8	28.5	28.8	8.3	23.8	24.1	9.4	20.3
pers	MNAR	50	19.6	23.6	63.0	60.0	18.2	55.5	52.6	19.9	48.5
		85	29.4	34.3	94.2	90.6	26.2	90.8	86.3	28.6	82.7

maximum simulation standard error = 0.43

informative nonresponse. Of the two stochastic household imputed values, the SL+ imputed values are slightly more accurate for ICC and MCAR and MAR, while the ML BLUP is the better of the two under the informative response mechanisms more typically associated with income nonresponse. After the log transformation the improvement in the ICC by introducing the ML BLUP is similar in magnitude to that seen on untransformed data.

The relative bias of the variance estimate for hourly wage rate is assessed in Table 4.5. To revisit the findings of Chapter 3, none of the deterministic

Table 4.5: Relative bias (%)of estimated variance - imputation using single stochastic BLUPs compared to deterministic BLUPs

NR model		ρ(%)	Deterministic				Stochastic			Stochastic log	
			Respmean	SL	SL+	ML	SL	SL+	ML	SL	ML
hh MCAR pers MCAR		20	-25.0	-23.4	-22.5	-22.6	0.2	0.3	0.3	-3.7	-0.7
		50	-25.0	-23.0	-17.8	-17.6	0.2	0.3	1.0	-3.3	1.7
		85	-25.1	-23.4	-6.9	-7.0	0.2	0.0	2.1	-4.0	7.2
hh MCAR pers MAR		20	-23.8	-21.6	-20.1	-20.7	1.9	4.9	4.3	-2.9	2.9
		50	-24.5	-21.9	-15.4	-15.2	1.7	-2.3	3.2	-2.8	2.3
		85	-26.8	-24.8	-8.9	-8.0	-3.4	-2.9	-0.4	-7.4	4.2
hh MCAR pers MNAR		20	-37.0	-35.6	-34.6	-34.6	-15.7	-18.0	-18.0	-17.2	-15.5
		50	-36.7	-34.9	-29.4	-29.4	-15.2	-17.0	-16.5	-16.5	-12.0
		85	-35.9	-34.3	-12.9	-14.4	-14.0	-8.3	-8.2	-16.3	-0.2
hh MNAR pers MCAR		20	-31.0	-29.5	-28.4	-28.3	-7.8	-11.6	-11.5	-10.6	-8.9
		50	-32.8	-30.9	-24.8	-24.8	-10.1	-12.3	-11.8	-12.2	-7.2
		85	-34.4	-32.8	-10.6	-12.1	-12.3	-6.0	-6.0	-14.8	2.6
hh MNAR pers MNAR		20	-40.6	-39.2	-38.2	-38.0	-20.6	-25.8	-25.7	-21.6	-21.0
		50	-37.9	-36.2	-30.0	-30.0	-16.8	-20.5	-20.2	-18.2	-13.8
		85	-37.7	-36.2	-10.9	-12.8	-16.5	-7.6	-8.5	-18.8	1.7

maximum simulation standard error = 0.52

BLUP imputed values resulted in accurate variance estimates, however there was some improvement from the use of household data in SL+ and ML for the highest level of ρ. Table 4.5 show that the stochastic BLUPs introduce an additional source of variation in the imputed values which results in much better variance estimates. Under MCAR, both of the SL BLUPs and the ML stochastic BLUP estimate the variance quite accurately. Under MNAR non-response mechanisms the stochastic BLUPs are again an improvement over the deterministic BLUPs. The household imputed values are slightly poorer

than the stochastic SL BLUP imputed values for low or moderate ρ, resulting in slightly higher bias in the variance. But when $\rho = 85\%$ the variance estimates using the household imputation methods are getting the variance about right, although still resulting in an underestimate of variance (at most 8.5%). The main benefit from the log transform is seen in the variance estimates under MNAR, where the stochastic ML BLUP with log transform results in some improvement in variance estimates under moderate levels of clustering, and large improvements under the highest level of clustering.

4.4.2 Multiple imputed values compared to single stochastic imputed values

This section includes a performance comparison of the single stochastic imputed values with multiple imputed values based on the single and multilevel models. Both raw and log transformed results are presented. Table 4.6 shows the predictive accuracy for imputed values as measured by RRMSE.

The use of multiple imputed values has resulted in improved accuracy, with the average over the $m = 30$ imputed values resulting in more precise imputed values. As with the stochastic and deterministic imputed values, the benefit from using ML rather than SL is greatest under high levels of clustering. MI results in similar accuracy levels for single-level imputed values

Table 4.6: RRMSE (%) - MI compared to single stochastic imputation using SL and ML BLUPs

NR model	ρ(%)	Single Stochastic		MI		MI log	
		SL	ML	SL	ML	SL	ML
hh MCAR pers MCAR	20	78.9	77.7	56.7	55.8	57.2	57.5
	50	78.2	70.3	56.0	56.5	56.6	52.6
	85	80.3	50.7	57.6	36.7	58.2	41.9
hh MCAR pers MAR	20	80.2	81.4	56.8	56.4	57.4	58.7
	50	79.5	72.1	56.0	52.1	56.8	53.9
	85	83.2	52.2	61.9	40.2	62.2	45.7
hh MCAR pers MNAR	20	80.5	77.3	65.0	63.9	64.6	63.8
	50	79.6	68.6	64.0	57.0	63.6	57.4
	85	81.4	47.7	65.0	39.4	64.6	43.5
hh MNAR pers MCAR	20	79.4	74.9	61.0	60.0	60.9	60.4
	50	78.7	67.0	61.4	54.6	61.2	55.4
	85	80.8	47.0	63.9	38.5	63.6	43.2
hh MNAR pers MNAR	20	79.9	74.6	66.4	65.2	65.6	64.6
	50	77.6	64.4	62.8	55.7	62.0	55.4
	85	79.0	43.7	64.0	37.8	63.1	41.8

maximum simulation standard error = 0.34

regardless of whether transformed data are used, and this is also reflected in the multilevel imputed values when $\rho = 20\%$ or 50%. Under the highest level of clustering however, the ML MI imputed values are slightly more accurate when using raw rather than log transformed data.

Table 4.7 shows the relative bias of the imputed values under both single and multiple imputed values. There is no major changes to the bias resulting from multiple rather than single imputed values. The previous issues with bias resulting from the ML BLUP under a log transform still exist under MI. The results for relative bias of the estimated mean reflect the findings for the

Table 4.7: Relative bias (%) - MI compared to single stochastic imputation using SL and ML BLUPs

NR model	ρ(%)	Single Stochastic		MI		MI log	
		SL	ML	SL	ML	SL	ML
hh MCAR pers MCAR	20	0.5	0.3	0.3	3.1	9.4	14.6
	50	0.4	0.1	0.2	0.1	9.3	15.1
	85	0.3	-0.0	1.9	0.0	9.3	16.3
hh MCAR pers MAR	20	-0.1	0.1	0.1	0.1	9.8	16.9
	50	0.6	-0.0	0.7	-0.1	10.3	15.3
	85	0.1	-1.2	1.0	-1.3	8.5	13.8
hh MCAR pers MNAR	20	-7.1	-6.5	-7.0	-6.5	1.3	5.5
	50	-7.2	-5.3	-7.1	-5.3	1.2	7.1
	85	-7.3	-2.6	-7.2	-2.6	1.1	10.9
hh MNAR pers MCAR	20	-3.8	-3.0	-4.0	-3.2	4.6	9.1
	50	-5.1	-3.0	-5.2	-3.1	3.2	9.6
	85	-6.5	-1.6	-6.6	-1.7	1.8	12.0
hh MNAR pers MNAR	20	-10.3	-9.1	-10.4	-9.2	-2.6	1.1
	50	-10.4	-7.1	-10.5	-7.1	-2.7	3.7
	85	-11.3	-3.1	-11.4	-3.1	-3.7	9.0

maximum simulation standard error = 0.22

bias of the imputed values (Table 7.3 in Appendix B).

The introduction of multiple imputed values leads to a large improvement in estimation of ICC under a single-level BLUP, and a small improvement for the ML BLUP, as shown in Table 4.8. The best imputation method for ICC is ML MI method, which reproduces the ICC well across all non-response mechanisms and levels of clustering, even under MNAR. MI on its own without the ML component is unable to achieve this, underestimating the ICC across all scenarios. The MI imputed values under a log transform give similar results for ICC as on the raw data, whether based on a SL or

Table 4.8: Expected value of estimated ICC - MI compared to single stochastic imputation using SL and ML BLUPs

NR model	ρ(%)	Single Stochastic		MI		MI log	
		SL	ML	SL	ML	SL	ML
hh MCAR pers MCAR	20	10.4	18.9	17.0	19.5	17.1	19.4
	50	27.4	47.3	35.3	46.4	35.7	46.4
	85	45.0	79.9	61.9	81.1	62.8	80.8
hh MCAR pers MAR	20	12.8	19.2	17.7	19.6	17.5	19.2
	50	30.3	51.0	37.0	48.5	37.6	48.5
	85	46.5	81.1	62.7	81.7	63.7	81.6
hh MCAR pers MNAR	20	9.5	21.0	16.8	20.1	16.9	19.6
	50	22.1	49.6	32.5	47.7	33.0	46.5
	85	33.1	83.0	55.2	82.8	56.3	81.4
hh MNAR pers MCAR	20	9.0	21.6	16.5	20.2	16.8	19.7
	50	21.3	50.3	32.0	48.1	32.7	46.8
	85	32.8	83.3	55.4	83.0	56.2	81.6
hh MNAR pers MNAR	20	8.3	24.1	16.4	20.9	16.7	19.8
	50	18.2	52.6	30.1	49.3	31.1	47.0
	85	26.2	86.3	51.4	84.6	52.7	82.6

ML model.

The ratios between the MI estimate of imputation variance to the actual imputation variance, var_{MI}/var_{true}, are shown in Table 4.9. MI generally results in under-estimation of the imputation variance. While the ML multiple imputed values have previously been shown to result in an improvement to RRMSE and ICC over SL models, the imputation variance estimates are poorer under ML than SL. This effect is seen for both the untransformed and transformed data and is worst under informative nonresponse and the highest levels of clustering where the imputation variance under MI is around

Table 4.9: Ratio of imputation variance under MI to true imputation variance under SL BLUP and ML BLUP methods

NR model	ρ(%)	MI		MI log	
		SL BLUP	ML BLUP	SL BLUP	ML BLUP
hh MCAR pers MCAR	20	69.2	70.7	52.8	45.2
	50	89.6	79.0	68.5	47.6
	85	114.0	88.0	84.9	45.5
hh MCAR pers MAR	20	57.7	62.1	50.6	46.6
	50	64.8	53.1	54.2	39.7
	85	63.1	32.5	66.7	31.9
hh MCAR pers MNAR	20	50.6	41.4	42.4	34.0
	50	66.7	45.6	55.4	42.8
	85	96.8	42.6	78.4	35.6
hh MNAR pers MCAR	20	58.3	43.3	44.8	35.2
	50	69.5	44.2	57.0	43.4
	85	90.2	42.3	74.3	36.3
hh MNAR pers MNAR	20	55.6	34.4	46.0	36.3
	50	65.2	31.4	52.6	37.4
	85	106.0	30.1	83.6	35.3

30−45% of the true imputation variance. This could potentially be addressed by an increase in the number of multiple imputed values, although 30 is already a relatively large number for multiple imputation. It is worth noting that unless MI is used, the imputation variance would usually be ignored altogether, so even an imperfect measure is better than nothing. Also, the total variance is made up of both sampling variance and imputation variance. MI is designed to help with the latter, but sampling variance would usually be at least as important in most surveys. Hence the biases shown in Table 4.9 would be small compared to the total variance.

4.4.3 Proportion of people on or below Federal Minimum Wage using deterministic and single stochastic imputation methods

Table 4.10: Various estimates of the percentage of adults on or below FMW - stochastic methods

NR model		ρ(%)	Actual	Deterministic SL	Deterministic ML	Stochastic SL	Stochastic ML	Stochastic log SL	Stochastic log ML
hh	MCAR	20	7.7	5.8	5.8	11.0	11.0	8.3	7.9
pers	MCAR	50	6.8	5.1	5.7	10.2	10.1	7.5	7.1
		85	6.7	5.0	7.2	10.2	9.7	7.4	6.8
hh	MCAR	20	7.7	5.8	5.8	11.0	11.2	8.4	8.1
pers	MAR	50	6.8	5.0	6.0	10.2	10.1	7.5	7.2
		85	6.7	4.8	7.5	9.7	9.4	7.3	6.9
hh	MCAR	20	7.7	6.0	6.0	11.0	10.5	8.6	8.1
pers	MNAR	50	6.8	5.3	5.7	10.2	9.4	7.8	7.3
		85	6.7	5.1	7.0	10.3	9.0	7.7	6.9
hh	MNAR	20	7.7	5.9	5.9	11.1	10.4	8.4	7.9
pers	MCAR	50	6.8	5.2	5.7	10.2	9.3	7.7	7.2
		85	6.7	5.1	7.0	10.2	8.9	7.6	6.9
hh	MNAR	20	7.7	6.1	6.1	11.0	9.9	8.7	8.1
pers	MNAR	50	6.8	5.4	5.8	10.3	8.8	7.9	7.2
		85	6.7	5.3	6.9	10.5	8.4	7.8	6.9

Table 4.10 shows that the stochastic imputed values result in quite different estimates for the proportion of people on or below FMW than the deterministic methods. Deterministic imputed values were previously shown to underestimate the percentage of adults on or below FMW, that is not imputing enough nonrespondents to lower incomes, but the stochastic imputed values over-estimate the true percentage, imputing far too many values in the

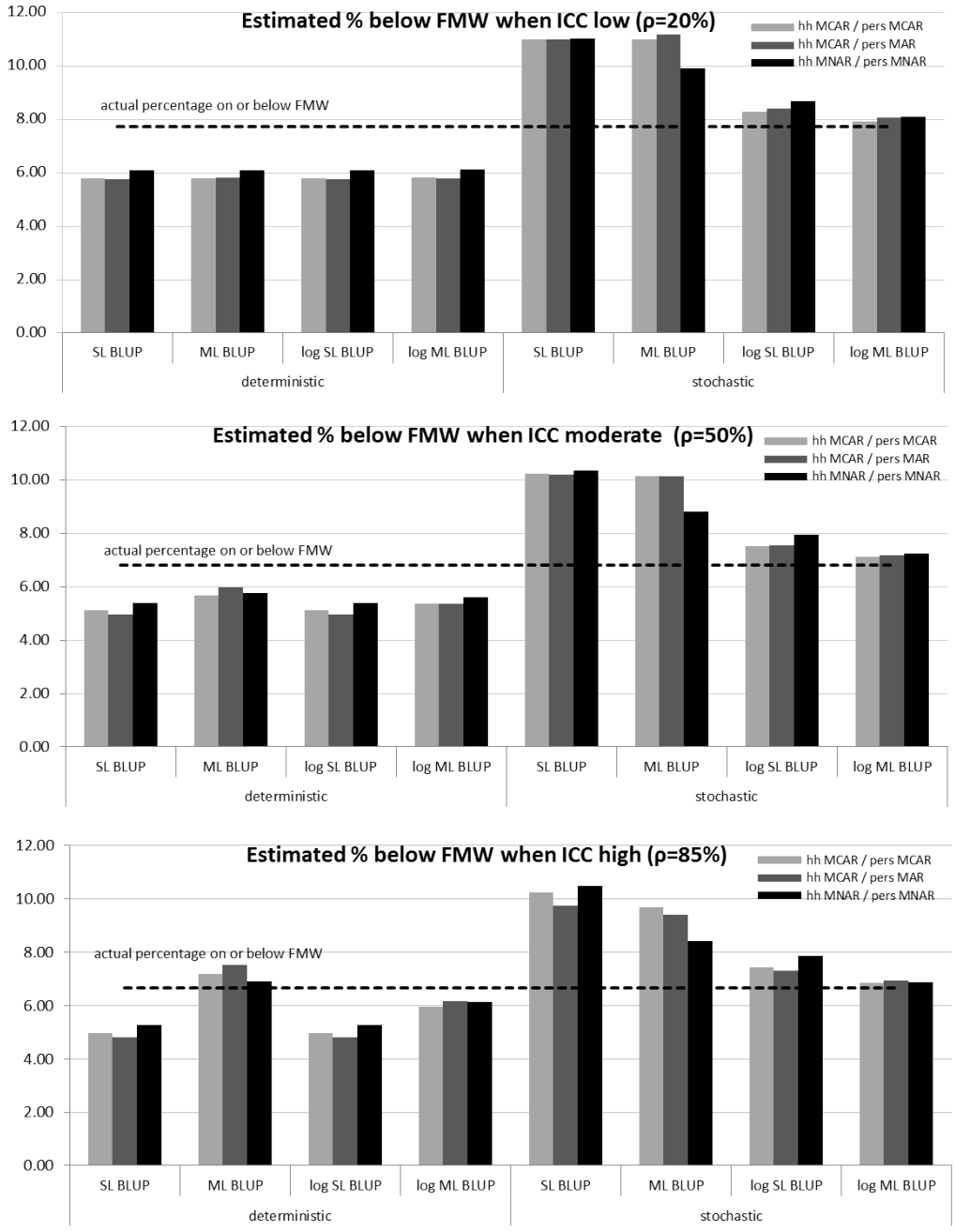

Figure 4.1: Simulation study - estimated percentage on or below Federal Minimum wage with stochastic imputed values

lower tail. In this specific example the log transform works well for both SL and ML imputed values, with the stochastic imputed values after log transform closely reproducing the true percentage of people on or below FMW. The ML BLUP stochastic imputed values with the log transformed data resulted in the best estimate of this proportion. However, this particular estimate is one of many that could be considered, and the ML BLUP was found to have bias issues with the log transformation, even after bias correction, which is cause for being cautious with this as an imputation method.

These results are also displayed in Figure 4.1, which shows the estimated percentage on or below federal minimum wage when $\rho = 20\%$, $\rho = 50\%$ and $\rho = 85\%$ respectively, for three of the nonresponse mechanisms.

4.5 Summary of Chapter 4

This chapter considered whether deficiencies in the deterministic ML imputation method, in particular over-estimation of ICC and under-estimation of variance, could be addressed by the use of a single stochastic or multiple imputed values.

The SL BLUP imputed values were poor for both ICC and variance estimation when either a single stochastic impute or multiple imputed values were used.

The stochastic ML BLUP was the standout imputation method for reproducing ICC. While the SL stochastic BLUP impute was very poor, with ICC underestimated by more than the deterministic SL BLUP, the stochastic ML BLUP reproduced the true ρ consistently well across all NR models and ICC levels.

The variance estimates under stochastic ML imputation were much improved under MNAR but there was still under-estimation of variance. This particular problem was not resolved by the introduction of multiple imputed values. The stochastic ML BLUP had significantly improved variance when ρ was high.

The application of stochastic imputation methods when estimating the proportion of people on or below federal minimum wage showed that the log-transformed multilevel stochastic imputed values do well in the lower tail of the income distribution, doing better than the deterministic and stochastic imputed values, particularly when using the multilevel model.

Chapter 5

Imputation of Binary Data using Generalized Linear Models and Generalized Linear Mixed Models

5.1 Introduction

In the previous two chapters imputation methods have been investigated for continuous variables using BLUPs under both a linear and linear mixed model. Household surveys generally contain a large number of data items

which are categorical or binary in nature, such as marital status, language, education, employment and disability. The proportion of people with income below FMW and the presence of important health condition such as diabetes and health risk factors such as smoking are also examples of key binary variables. In these cases the previous models based on a continuous distribution are not applicable.

The focus of this chapter is the development of new methods appropriate for dealing with missing data for binary variables in a household survey. Imputation models will be developed based on logistic and probit models with and without a random household intercept. The proposed imputation methods will be assessed against existing imputation methods by a simulation study using the British Household Panel Survey (BHPS). Both the single and multilevel logit and probit models are used to generate a single stochastic impute, and then multiple imputed values. Evaluation of the imputation methods is carried out by assessing various qualities of these imputed values across several potential nonresponse mechanisms over a set of replicates.

Section 5.2 has a description of how imputed values are derived under the logit and probit models using both Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs). In Section 5.3 the simulation study using two binary variables from the BHPS is described. The results and findings are detailed in Section 5.4. The chapter concludes with a summary

of the value of introducing the household structure into an imputation model for binary variables.

5.2 Imputation methods

5.2.1 Generalized Linear and Generalized Linear Mixed Model with logit link function

The first imputation method makes use of the logit link function. Using a Generalized Linear Model (GLM), Y_{ij} is assumed to have a binary distribution with the logit function used to link the linear predictor $\mathbf{x}_{ij}\boldsymbol{\beta}$ and Y_{ij}:

$$P(Y_{ij} = 1) = f(\mathbf{x}_{ij}^T\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})} \tag{5.1}$$

Similarly when using a Generalized Linear Mixed Model (GLMM), Y_{ij} is assumed to have a binary distribution, but now the logit function is used to link the linear predictor $\mathbf{x}_{ij}\boldsymbol{\beta}$, the random household effect u_j and Y_{ij}:

$$P(Y_{ij} = 1|u_j) = g(\mathbf{x}_{ij}^T\boldsymbol{\beta} + u_j) = \frac{\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + u_j)}{1 + \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + u_j)} \tag{5.2}$$

where $u_j \sim N(0, \sigma_u^2)$.

The derivation of imputed values under (5.1) is straightforward. The logit model is first fit to the respondents' data. We then estimate $P(Y_{ij} = 1)$ by

replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$. and use this probability to generate Bernoulli random variables, resulting in binary imputed values. The underlying assumption of this model is that nonresponse in either MCAR or MAR, therefore there may be nonresponse bias unaccounted for when nonresponse is generated under the MNAR assumption. The mixed model of (5.2) is more complex for derivation of imputed values as it includes the unknown random household effect.

To generate imputed values under (5.2), suppose persons $1, ..., n_{rj}$ in household j respond, and person n_{rj+1} is a nonrespondent. The best impute for missing $Y_{n_{rj+1},j}$ is $E[Y_{n_{rj+1},j}|Y_{1j}, ..., Y_{n_{rj},j}]$, which although this is not binary, can be used to generate a Bernoulli random variable as a binary impute. This is not straightforward to calculate as it requires integration over the values of u_j, and also depends on $\boldsymbol{\beta}$ and σ_u^2. The empirical Bayes estimator can be calculated by substituting estimates for $\boldsymbol{\beta}$ and σ_u^2. To avoid numerical integration over the u_j, a stochastic method can be used:

$$E[Y_{n_{rj+1},j}|Y_{1,j}, ..., Y_{n_{rj+1},j}] = E[g(\mathbf{x}_{n_{rj+1},j}\boldsymbol{\beta} + u_j)|Y_{1j}, ..., Y_{n_{rj,j}}] \tag{5.3}$$

$$= E_{u^*}[g(\mathbf{x}_{n_{rj+1,j}}\boldsymbol{\beta} + u_j^*)] \tag{5.4}$$

where u_j^* are drawn from the the distribution of u_j conditional on the observed values $Y_{1j}, ..., Y_{n_{rj,j}}$.

Firstly model (5.2) is fitted to fully responding households, resulting in estimates for the unknown parameters σ_u^2 and $\boldsymbol{\beta}$. This model was fit using maximum likelihood with quadrature approximation in SAS PROC GLIMMIX. A set of K^* independent values for each household j containing nonrespondents, $u_{jk}^* : \quad k = 1, ..., K^*$, are then generated from $N(0, \hat{\sigma}_u^2)$. These can then be used to generate Y_{ijk}^* using:

$$P(Y_{ijk}^* = 1 | u_{jk}^*, Y_{1j}) = g(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_{jk}^*) \tag{5.5}$$

for $i = 1, ..., n_{rj}$ where Y_{1j} is a respondent value in household j.

The probability in (5.5) cannot be used to create an impute because u_{jk}^* are generated from the estimated marginal distribution of u_j ($u_j \sim N(0, \hat{\sigma}_u^2)$), but we require draws from the distribution of u_j given $Y_{1j}, ..., Y_{n_rj}$. To achieve this a reduced set of K^{**} replicates are now defined by only retaining u_{jk}^* from replicates where the generated values of Y_{ijk}^* are equivalent to the observed values Y_{ij}. By this device, the reduced set of random effects, which is written as $u_{jk}^* : \quad k = 1, ..., K^{**}$, are draws from the distribution of u_j given the observed data. For example, in the special case of two people per household with one respondent, u_{jk}^* is retained when $Y_{1jk}^* = Y_{1j}$. This reduced set of household random effects is notated using $u_{jk}^* : k = 1, ..., K^{**}$.

Multiple imputed values for nonrespondent $Y_{n_{rj+1},j}$ can be generated us-

ing:

$$P(Y^{**}_{n_{rj+1},jk} = 1|u^{**}_{jk}) = g(\mathbf{x}_{n_{rj+1},jk}\hat{\boldsymbol{\beta}} + u^{**}_{jk}) \tag{5.6}$$

To choose a single impute just one of these imputed values could be used, or the mean of a set of imputed values (although this in general would not be 0 or 1):

$$\hat{Y}_{n_{rj+1},j} = \frac{1}{K^{**}} \sum_{k=1}^{K^{**}} Y^{**}_{n_{rj+1},jk} \tag{5.7}$$

5.2.2 Generalized Linear and Generalized Linear Mixed Model with probit link function

The second imputation method using a probit model will be derived for first a GLM then a GLMM.

Let $Z_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$ be an underlying, unobserved response variable with $\epsilon_{ij} \sim N(0, 1)$ such that:

$$Y_{ij} = \begin{cases} 1, & \text{if } Z_{ij} \geq 0 \\ 0, & \text{if } Z_{ij} < 0 \end{cases}$$

$\mathbf{Z_j} = (Z_{1j}, ..., Z_{n_j,j})$ are independent multivariate normal random variables:

$$E[Z_{ij}] = \mathbf{x}_{ij}^T \boldsymbol{\beta}$$

$$Var(Z_{ij}) = 1$$

$$Cov(Z_{ij}, Z_{i'j}) = 0$$

Then

$$\begin{aligned} P(Y_{ij} = 1) &= P(Z_{ij} \geq 0) \\ &= P(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij} \geq 0) \\ &= P(\epsilon_{ij} < \mathbf{x}_{ij}^T \boldsymbol{\beta}) \qquad \text{(by symmetry)} \\ &= \Phi(\mathbf{x}_{ij}^T \boldsymbol{\beta}) \end{aligned}$$

where Φ is the cumulative distribution function for the standard normal distribution.

The probit model can be fit to the responding data to generate imputed values for missing Y_{ij}. As in Section (5.2.1) a Bernoulli random variable can be used to generate binary values if required.

Correspondingly the GLMM is specified by defining an underlying, un-

observed response variable $Z_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_j + \epsilon_{ij}$ such that:

$$Y_{ij} = \begin{cases} 1, & \text{if } Z_{ij} \geq 0 \\ 0, & \text{if } Z_{ij} < 0 \ . \end{cases}$$

with $u_j \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0,1)$. $\mathbf{Z_j} = (Z_{1j}, ..., Z_{n_j,j})$ are independent multivariate normal random variables with:

$$E[Z_{ij}] = \mu_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}$$

$$Var(Z_{ij}) = 1$$

$$Cov(Z_{ij}, Z_{i'j}) = \rho \qquad (i \neq i').$$

As in Section (5.2.1) an impute under this model requires a draw from Y_{2j} given Y_{1j}, in the case of households of size 2 with one nonrespondent (without loss of generality person 1 is assumed the respondent).

If Z_{2j} is known then Y_{2j} follows, therefore draws are first generated from the distribution of Z_{2j} given Y_{1j}, or equivalently, Z_{2j} given sign(Z_{1j}).

The distribution of Z_{2j} conditional on Z_{1j} comes from the known properties of the multivariate normal distribution (also see Section 4.2):

$$Z_{2j}|Z_{1j} \sim N(\mu_{2j} + \rho(Z_{1j} - \mu_{1j}), 1 - \rho^2) \tag{5.8}$$

The first step is to fit the GLMM probit model to data from responding

households, resulting in estimates for the unknown parameters σ_u^2, $\boldsymbol{\beta}$ and ρ.

To generate Z_{2j} also requires Z_{1j}. Now $Z_{1j} \sim N(\mathbf{x}_{1j}\boldsymbol{\beta}, 1)$, but Y_{1j} is known so can be used when generating Z_{1j}. Let Z_{1j}^* be draws from the distribution of $Z_{1j}|Y_{1j}$. When $Y_{1j} = 1$ then $Z_{1j} > 0$, and conversely when $Y_{1j} = 0$ then $Z_{1j} < 0$. Therefore the distribution of $Z_{1j}|Y_{1j}$ is *truncated normal*:

$$Z_{1j} \sim TrN(x_{1j}\boldsymbol{\beta}, 1) \text{ for } 0 < Z_{1j} \text{ conditional on } Y_{1j} = 1 \text{ and} \tag{5.9}$$

$$Z_{1j} \sim TrN(x_{1j}\boldsymbol{\beta}, 1) \text{ for } 0 > Z_{1j} \text{ conditional on } Y_{1j} = 0 \tag{5.10}$$

A truncated normal distribution results from a normally distributed random variable which is bounded below, above, or both. Formally, if $Z \sim N(\mu, \sigma^2)$ and Z is constrained to take a value between a and b, then Z conditional on $a < Z < b$ has the truncated normal distribution with pdf:

$$f(Z; \mu; a, b) = \frac{\frac{1}{\sigma}\phi(\frac{Z-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \tag{5.11}$$

where ϕ represents the standard normal density function and Φ represents the corresponding cumulative distribution function.

Values for Z_{1j}^*, are therefore generated by taking random draws from the truncated normal distributions above using the respondents value Y_{1j} and $\mathbf{x}_{1j}$ and the estimate for $\boldsymbol{\beta}$ from responding households.

Following this a value for Z_{2j}, Z^*_{2j} can be drawn for nonrespondents, using (5.8), Z_{1j} and $\hat{\rho}$. To convert these to binary values, $Y^*_{2j} = 1$ when $Z^*_{2j} > 0$ and $Y^*_{2j} = 0$ when $Z^*_{2j} < 0$. It is also possible to generate a continuous impute for Y_{2j} by taking the mean over multiple binary imputed values, $Y^*_{2j} = \sum_{k=1}^{K} Y^*_{2jk}$. However, in some cases a binary impute is preferable to be consistent with the respondent data.

Because of the difficulty of generating stochastic imputed values for binary variables some practitioners use naive methods, such as assuming the response variable is normally distributed and employing a rounding method. The pitfalls of this approach are demonstrated in Horton, Lipsitz, and Parzen (2003). These pitfalls are avoided using the methods above, which generate binary imputed values, therefore avoiding the need for rounding.

5.3 Simulation study

The simulation study for binary data will make use of a longitudinal survey, the British Household Panel Survey (BHPS), which has been running since 1991. The survey is run by the Economic and Social Research Council (ESRC) UK Longitudinal Studies Centre with the Institute for Social and Economic Research (ISER) at the University of Essex. The BHPS is a multipurpose study which is household based, where all adult members aged 16

and older are interviewed, from a panel of 5,500 households resulting in approximately 10,300 individual respondents in wave 1 (Taylor et al., 2010). It includes questions from a range of topics to improve understanding of social and economic factors at person and household level in Britain. A single wave of data, from the 15th wave was chosen for the simulation study. This wave consisted of a sample of 8,703 households containing 15,627 individuals, collected between September 2005 and May 2006. As in the previous simulation study the simulation was restricted to households with two respondents.

5.3.1 Imputation variables from BHPS

Two variables were elected from this study to represent different prevalence rates and levels of clustering. The first variable was voting intention, derived using the variable *Party which would vote for tomorrow*. This question was asked of surveys participants who supported a particular political party or were closer to one party than the other. A binary variable was created representing whether the person would vote labour which had overall prevalance 25% in those that responded and identified a party. There were 542 people from 271 households of size two where both persons responded to this question with a valid response (those who identified 'none' were excluded).

The second variable chosen was labour force status, derived using the

Table 5.1: BHPS simulation variables

Variable	No. hh	No. pers	$\hat{p}$	$\hat{p}_{same}$
vote labour	271	542	0.25	0.80
employed	4,031	8,062	0.61	0.73

variable *Current economic activity*. A binary variable was created representing those currently employed/ self-employed versus unemployed, students, retired and other categories. The prevalence was approximately 61%. There were 8,062 people from 4,031 households of size two where both persons responded. The proportion of people with the same response in a household was also considered and there was moderate disagreement within-household as shown in Table 5.1 under the column titled $\hat{p}_{same}$.

5.3.2 Simulating nonresponse

The simulation of nonresponse in the BHPS closely followed the method used for the simulation from the HILDA survey described in Sections 3.3.2 and 4.3. The fully observed component of the sample was used to generate $K = 250$ simulated samples with item nonresponse, to isolate the impact of the item nonresponse mechanism and imputation method as distinct from population or sample variation. Approximately half of households were designated to have item nonresponse and one of the two people within each nonresponding household was selected to be a item nonrespondent according to the differ-

ent response models described below. The resulting item response rate was approximately 75% under each scenario. Five alternative models were used to generate nonresponse. The first has data Missing Completely At Random (MCAR), the second Missing at Random (MAR) and the others Not Missing At Random (MNAR). These nonresponse mechanisms are described below. They are similar to chapters 3 and 4 with small modifications accounting for the binary nature of the variables and the specific variables chosen.

As in chapters 3 and 4, p_{1j} and p_{2j} are the probabilities of response for person 1 and person 2 in household j respectively. The probabilities will now specified for each nonresponse scenario.

Households MCAR and persons MCAR: Same as model employed in Chapters 3 and 4 as described in Section 3.3.2.

Households MCAR and persons MAR: Same as model employed in Chapters 3 and 4 as described in Section 3.3.2.

Households MCAR and persons MNAR: The probability of being a respondent was dependent on Y_{ij}, with labour voters having an odds ratio of 0.2:

$$p_{ij} = \frac{\exp(\beta_0 - 0.2Y_{ij})}{1 + \exp(\beta_0 - 0.2Y_{ij})}$$

Households MNAR and persons MCAR: Households are partially or fully responding, with the odds of the household falling in the first category de-

creasing by approximately 10% when there is a single labour voter, and approximately 20% with two labour voters. Within partially responding households, one person was randomly an item respondent, the other had item nonresponse.

$$p_{ij} = \frac{\exp(\beta_0 - 0.2\bar{Y}_j)}{1 + \exp(\beta_0 - 0.2\bar{Y}_j)}$$

where $\bar{Y}_j = (Y_{1j} + Y_{2j})/2$.

Households MNAR and persons MNAR: Lastly, households and persons were assigned to be partially or fully responding, both MNAR. The probability of the household falling in the first category decreased by 10% with each additional labour voter in the household, and each persons probability of response also decreased by 20% if they voted labour. This implies that person probabilities of response depended on how many vote labour in the household:

$$p_{ij} = \frac{\exp(\beta_0 - 0.2\bar{Y}_j - 0.2Y_{ij})}{1 + \exp(\beta_0 - 0.2\bar{Y}_j - 0.2Y_{ij}))}$$

In all five scenarios, β_0 was calculated such that the average of p_{ij} was exactly or approximately 0.75, so that approximately 75% of people were item respondents, leading to 50% of households having full response and 50% of households having one item nonrespondent. The probability of each

possible household response was calculated as in Section 3.3.2

5.4 Results

The analysis is carried out on each binary response variable, whether employed, and voting intention of labour.

5.4.1 Single imputation methods

The RRMSE and Relative Bias resulting from each imputation method are shown below in Table 5.2, for the two response variables Y_1 (labour), and Y_2 (employed). The imputation methods compared in this section are single stochastic imputation methods, firstly using a generalized linear model with logit link (single-level) and then a generalized linear mixed model with logit link (multilevel model). The latter two columns show these same imputation methods under the imputation model using a probit link.

The RRMSE values are higher than seen for the continuous variable in the previous two chapters, which naturally follows from imputing a 0/1 variable. The residuals are always either 0, 1 or -1, which imputes large errors relative to the prevalence.

Under both logit and probit models there was a significant improvement in the accuracy of the imputed values of voting intention of labour when

Table 5.2: RRMSE (%) - single imputation methods for Y_1 - labour and Y_2 - employed

NR model		logit		probit	
hh	pers	SL	ML	SL	ML
vote labour					
MCAR	MCAR	242·2	218·0	242·5	217·8
MCAR	MAR	232·6	206·7	229·9	207·4
MCAR	MNAR	219·0	198·8	220·0	199·5
MNAR	MCAR	224·5	202·6	223·8	203·6
MNAR	MNAR	205·9	186·4	203·4	188·4
employed					
MCAR	MCAR	103·0	101·1	104·4	101·4
MCAR	MAR	96·2	94·4	96·7	94·9
MCAR	MNAR	97·1	95·4	98·5	95·7
MNAR	MCAR	98·8	97·2	100·3	97·5
MNAR	MNAR	93·4	92·3	95·6	92·6

using the GLMM compared to the GLM. The multilevel imputed values averaged about 10% lower RRMSE than the single-level BLUP across all nonresponse mechanisms. For employment status, and under both logit and probit models, there is a much smaller reduction in RRMSE when using the GLMM compared to the GLM. For this variable the multilevel imputed values averaged about 2% smaller RRMSE than the single-level imputed values.

While there were gains in predictive accuracy in the imputed values arising under a generalized linear mixed model over a generalized linear model, there is no improvement made in the relative bias of the imputed values. The following table, Table 5.3 shows the relative biases of the imputed values for voting intention of labour and whether employed.

Table 5.3: Relative Bias (%) - single imputation methods for Y_1 - voting preference labour and Y_2 - employed

NR model		logit		probit	
hh	pers	SL	ML	SL	ML
vote labour					
MCAR	MCAR	0·9	0·5	1·4	2·6
MCAR	MAR	2·2	−0·8	−4·8	−0·6
MCAR	MNAR	−12·9	−11·9	−15·3	−9·9
MNAR	MCAR	−9·7	−6·7	−14·0	−7·6
MNAR	MNAR	−20·3	−19·2	−29·4	−18·9
employed					
MCAR	MCAR	0·0	−0·3	−1·2	−0·6
MCAR	MAR	−0·4	−0·5	−0·3	−1·2
MCAR	MNAR	−6·2	−7·2	−8·8	−7·6
MNAR	MCAR	−4·2	−5·2	−7·3	−5·5
MNAR	MNAR	−9·6	−11·3	−14·4	−11·7

For voting preference of labour, the multilevel logit imputed values have less relative bias than the single-level logit imputed values. Under a probit model this is generally also the case, except when nonresponse is MCAR. The multilevel logit model results in slightly larger bias than the single-level model across all nonresponse mechanism for whether employed, but there were gains when using the probit ML model over the single-level. Overall the multilevel imputed values are not resulting in any large gains for bias.

The next evaluation criteria assessed how well the imputation methods retained clustering. Table 5.4 shows the percentages of household members where the binary variable was the same for both households.

The mixed binary imputation models are advantageous when looking at

Table 5.4: Households with same voting intention of labour (%) and same employment status (%) - single imputation methods

NR model			logit		probit	
hh	pers	true	SL	ML	SL	ML
vote labour						
MCAR	MCAR	79·3	71·0	79·0	71·0	79·1
MCAR	MAR	79·3	71·0	80·3	71·4	80·1
MCAR	MNAR	79·3	71·1	79·1	71·4	79·0
MNAR	MCAR	79·3	71·2	79·2	71·2	78·9
MNAR	MNAR	79·3	71·3	79·1	71·8	79·0
employed						
MCAR	MCAR	72·2	66·2	70·3	65·7	70·5
MCAR	MAR	72·2	66·7	70·5	66·6	70·9
MCAR	MNAR	72·2	65·8	69·9	65·3	70·1
MNAR	MCAR	72·2	66·0	70·0	65·4	70·1
MNAR	MNAR	72·2	65·7	69·5	64·8	69·5

the percentage of households where the people in the household responded in the same way. The proportion matching is underestimated by around 10% under SL imputation models, but is estimated very accurately when using ML imputed values. For intention to vote labour, the relationship within households is maintained under the multilevel model, with both the probit and logit models, while the single-level imputed values result in too much difference between people within households. For employment status, the generalized linear mixed model imputed values which under-estimated clustering but were much closer to the true clustering than the generalized linear model imputed values.

The relative bias of the estimated population proportions are shown in

Table 5.5. The results reflect the findings for individual imputed values.

Table 5.5: Relative Bias of population proportion (%) - single imputation methods for Y_1 - voting preference labour

NR model		logit		probit	
hh	pers	SL	ML	SL	ML
vote labour					
MCAR	MCAR	−0·2	−0·3	−0·2	0·2
MCAR	MAR	0·2	−0·6	−1·7	−0·6
MCAR	MNAR	−3·9	−3·7	−4·8	−3·1
MNAR	MCAR	−3·1	−2·2	−3·9	−2·4
MNAR	MNAR	−6·4	−6·1	−9·3	−5·8
employed					
MCAR	MCAR	0·0	−0·1	−0·3	−0·2
MCAR	MAR	−0·1	−0·1	−0·1	−0·3
MCAR	MNAR	−1·6	−1·9	−2·3	−2·0
MNAR	MCAR	−1·1	−1·3	−1·9	−1·4
MNAR	MNAR	−2·6	−3·1	−3·9	−3·2

5.4.2 Multiple imputation methods

Predictive accuracy as measured by RRMSE and relative bias for each multiple imputation method is shown below. Table 5.6 shows RRMSEs for single stochastic and multiple imputed values.

Multilevel multiple imputed values result in improved RRMSE compared to single imputed values under both the logit and probit models and for both dependent variables. Multiple imputation results in 25-30% improvement in RRMSE compared to single imputation, and this is consistent for both logit and probit models and across all nonresponse mechanisms. The

Table 5.6: Relative Root Mean Square Error (%) - single and multiple imputation methods for Y_1 (labour) and Y_2 (employed)

NR model		logit				probit			
hh	pers	SL	ML	SL MI	ML MI	SL	ML	SL MI	ML MI
vote labour									
MCAR	MCAR	242·2	218·0	174·2	157·1	242·5	217·8	174·6	157·0
MCAR	MAR	232·6	206·7	166·5	153·0	229·9	207·4	167·2	152·7
MCAR	MNAR	219·0	198·8	161·3	145·5	220·0	199·5	162·6	145·7
MNAR	MCAR	224·5	202·6	164·8	148·0	223·8	203·6	165·4	147·9
MNAR	MNAR	205·9	186·4	154·3	138·8	203·4	188·4	155·8	139·0
employed									
MCAR	MCAR	103·0	101·1	73·2	71·4	104·4	101·4	73·9	71·6
MCAR	MAR	96·2	94·4	69·4	67·8	96·7	94·9	69·8	68·1
MCAR	MNAR	97·1	95·4	68·3	66·8	98·5	95·7	69·3	67·0
MNAR	MCAR	98·8	97·2	69·9	68·3	100·3	97·5	70·8	68·5
MNAR	MNAR	93·4	92·3	65·6	64·4	95·6	92·6	67·1	64·6

improvements from a generalized linear mixed imputation model rather than a generalized linear imputation model found for single imputed values also hold for multiple imputed values, that is an additional 10% accuracy improvement for voting intention and around 2% for employment status. The major reason for the lower RRMSEs under multiple imputation is likely that the impute is effectively averaged over many stochastic imputed values, hence non-integer imputed values are permitted.

The relative bias of imputed values was then calculated under each of the multiple imputation methods and is show in Table 5.7 for the two study variables.

The biases of imputed values under multiple imputation reflect the results

Table 5.7: Relative Bias (%) - single and multiple imputation methods for Y_1 (labour) and Y_2 (employed)

NR model		logit				probit			
hh	pers	SL	ML	SL MI	ML MI	SL	ML	SL MI	ML MI
vote labour									
MCAR	MCAR	0·9	0·5	2·0	1·0	1·4	2·6	2·5	1·0
MCAR	MAR	2·2	−0·8	1·4	−0·1	−4·8	−0·6	−3·2	−0·2
MCAR	MNAR	−12·9	−11·9	−12·3	−12·0	−15·3	−9·9	−15·8	−11·9
MNAR	MCAR	−9·7	−6·7	−8·6	−8·4	−14·0	−7·6	−13·5	−8·2
MNAR	MNAR	−20·3	−19·2	−20·5	−19·8	−29·4	−18·9	−28·7	−20·1
employed									
MCAR	MCAR	0·0	−0·3	−0·1	−0·4	−1·2	−0·6	−1·6	−0·8
MCAR	MAR	−0·4	−0·5	−0·4	−0·7	−0·3	−1·2	−0·3	−1·5
MCAR	MNAR	−6·2	−7·2	−6·2	−7·2	−8·8	−7·6	−8·8	−7·5
MNAR	MCAR	−4·2	−5·2	−4·1	−5·2	−7·3	−5·5	−7·4	−5·6
MNAR	MNAR	−9·6	−11·3	−9·6	−11·4	−14·4	−11·7	−14·3	−11·8

under single imputation. There is no additional gain from using MI rather than single imputed values.

The percentage of households with the same voting intention, and the same employment status is shown in Table 5.8 for each multiple imputation method.

The results show that the multiple imputed values retain the appropriate level of clustering as seen in the single imputed values for both response variables and under both probit and logit imputation models. Introducing multiple imputed values under a single-level model did not improve the estimates of the proportion of people with the same response. This was only achieved with the introduction of the multilevel imputation model.

Table 5.8: Households with same response (%) for Y_1 (labour) and Y_2 (employed) - single and multiple imputation methods

NR model			logit				probit			
hh	pers	true	SL	ML	SL MI	ML MI	SL	ML	SL MI	ML MI
vote labour										
MCAR	MCAR	79·3	71·0	79·0	71·0	79·2	71·0	79·1	71·0	79·2
MCAR	MAR	79·3	71·0	80·3	71·2	80·7	71·4	80·1	71·4	80·5
MCAR	MNAR	79·3	71·1	79·1	71·3	79·5	71·4	79·0	71·4	79·3
MNAR	MCAR	79·3	71·2	79·2	71·0	79·3	71·2	78·9	71·2	79·2
MNAR	MNAR	79·3	71·3	79·1	71·2	79·4	71·8	79·0	71·7	79·2
employed										
MCAR	MCAR	72·2	66·2	70·3	66·1	70·4	65·7	70·5	65·7	70·6
MCAR	MAR	72·2	66·7	70·5	66·7	70·5	66·6	70·9	66·6	70·9
MCAR	MNAR	72·2	65·8	69·9	65·8	70·0	65·3	70·1	65·3	70·2
MNAR	MCAR	72·2	66·0	70·0	66·0	70·1	65·4	70·1	65·4	70·2
MNAR	MNAR	72·2	65·7	69·5	65·6	69·5	64·8	69·5	64·9	69·7

Lastly the relative bias for the estimate of population proportion is shown for both response variables in Table 5.9. As with the bias of the imputed values, there is again no major improvements in bias for the estimated population proportions from the introduction of multiple rather than single imputed values.

5.4.3 Multiple Imputation variance

The results for ratio between variance under MI and true variance for labour voting preference and whether employed are shown in Table 5.10.

The ratio between imputation variance and true variance excludes the

Table 5.9: Relative Bias of population proportion (%) - single and multiple imputation methods for Y_1 (labour) and Y_2 (employed)

NR model		logit				probit			
hh	pers	SL	ML	SL MI	ML MI	SL	ML	SL MI	ML MI
vote labour									
MCAR	MCAR	−0·2	−0·3	0·0	−0·2	−0·2	0·2	−0·1	−0·3
MCAR	MAR	0·2	−0·6	−0·1	−0·4	−1·7	−0·6	−1·3	−0·4
MCAR	MNAR	−3·9	−3·7	−3·8	−3·7	−4·8	−3·1	−4·8	−3·7
MNAR	MCAR	−3·1	−2·2	−2·7	−2·6	−3·9	−2·4	−4·2	−2·6
MNAR	MNAR	−6·4	−6·1	−6·4	−6·2	−9·3	−5·8	−9·0	−6·3
employed									
MCAR	MCAR	0·0	−0·1	0·0	−0·1	−0·3	−0·2	−0·3	−0·2
MCAR	MAR	−0·1	−0·1	−0·1	−0·2	−0·1	−0·3	−0·1	−0·4
MCAR	MNAR	−1·6	−1·9	−1·6	−1·9	−2·3	−2·0	−2·3	−2·0
MNAR	MCAR	−1·1	−1·3	−1·1	−1·4	−1·9	−1·4	−1·9	−1·5
MNAR	MNAR	−2·6	−3·1	−2·6	−3·1	−3·9	−3·2	−3·9	−3·2

component of variance due to sampling. It is measuring the ability of MI to reproduce the variance due to imputation excluding sampling variability. For both response variables MI results in under-estimation of the variance, across all nonresponse mechanisms. The magnitude is not overly large, and would likely be dominated by the variation due to sampling. The ratio of imputation variance to true variance was fairly high, between 75% and 95% for both the SL and ML logit methods and the ML probit method. The SL probit imputed values were lower, between 50% and 75%.

Table 5.10: Ratio of imputation variance to true variance for population proportion (%) - for Y_1 (labour) and Y_2 (employed)

NR model		logit		probit	
hh	pers	SL MI	ML MI	SL MI	ML MI
vote labour					
MCAR	MCAR	89·4	78·2	52·7	80·5
MCAR	MAR	85·3	70·7	53·4	72·9
MCAR	MNAR	92·5	76·1	55·7	79·5
MNAR	MCAR	91·2	81·6	54·3	83·4
MNAR	MNAR	86·4	73·3	51·1	70·7
employed					
MCAR	MCAR	86·0	71·1	61·9	71·7
MCAR	MAR	95·3	77·5	72·1	82·9
MCAR	MNAR	92·1	80·0	72·6	81·3
MNAR	MCAR	92·6	79·5	71·2	81·5
MNAR	MNAR	93·7	80·7	75·2	81·8

5.5 Summary of Chapter 5

This chapter looked at several methods for imputing missing binary variables for household surveys. Imputed values generated under a generalized linear mixed model were compared to a generalized linear model using both a logit and probit link function. Single and multiple imputed values derived under these models were compared for two binary variables using a simulation from the BHPS dataset. These were assessed using a set of evaluation criteria appropriate for binary variables.

For single imputation methods there was around 10% improvement in accuracy when using a GLMM compared to the GLM impute. A further 25-

30% improvement in accuracy occurred by introduction of multiple imputed values. The use of mixed models had little improvement on the biases of the population proportion estimates.

With regards to clustering, the multilevel single imputation methods were clearly superior to their single-level counterparts, closely reproducing the proportion of households with the same value for both voting intention and employment status. For this reason there was no additional gain from the introduction of multiple imputed values. MI variance estimates were reasonably accurate when use with multilevel models having biases between -2.9% and 5%.

Chapter 6

Imputation of Continuous or Binary Data using Donor Methods

6.1 Introduction

The previous three chapters addressed imputation from a linear and generalized linear model perspective, for both continuous and binary variables. Single and multiple imputation methods were considered as well as imputation using mixed models. Imputation methods based on linear models have known weaknesses, particularly their poor results in reproducing variation due to the model predictions focussing on a mean. Multiple imputation im-

proves on variance estimation but still has limitations, such as distributional assumptions. For example the response distribution, or its transformation, is normal. A widely used alternative is to use donor methods, which are not constrained by concept of imputing a mean value. Advantages include imputing draws from the observed response distribution creating potentially more plausible imputed values, the potential to impute many variables simultaneously without need for a multivariate model, and being able to handle a mixture of variable types (e.g. continuous, positive continuous, binary, discrete, ordinal) without a need to create an explicit model. While donor methods are associated with improved variance estimation properties, their ability to reproduce within-household clustering is unclear.

The goal of this chapter is to conduct a comprehensive investigation of donor methods to determine which are most appropriate in the household survey setting. New methods are proposed which make use of information on household respondents. These have been evaluated side-by-side with existing methods to determine relative strengths and weaknesses.

Donor methods create an impute from a random draw from the set of respondents, resulting in a continuous or binary impute as required. The method of selecting a donor may be as simple as a random hot deck, which replaces the missing value for a nonrespondent, the recipient, by a respondents variable (e.g. Kalton and Kasprzyk, 1982), or as with linear models,

make use of auxiliary variables for which data is available on both respondent and nonrespondents. Within-class donor imputation uses categorical variables to create imputation classes from which a donor is drawn. For categorical auxiliary variables this is straightforward. Continuous auxiliary variables may be formed into categories. An alternative for continuous variables is a donor method such as nearest neighbour (Chen and Shao, 2000) which finds a donor with minimum distance from the recipients continuous auxiliary variable. This approach may also be carried out within imputation classes formed by other auxiliary variables.

One constraint of traditional donor methods is they often focus on the attributes of the individual without accounting for information about other responding members of the household. Bankier (1999) describes a household multivariate donor imputation method used for a small number of variables (age, sex, marital status, common-law status and relationship) in the 1996 Canadian Census. This method, known as NIM (New Imputation Methodology) is a combined editing and imputation strategy based on the principle of minimum change (Fellegi and Holt, 1976). A set of variables are imputed using a donor that ensures a record meets all edit rules, while also making the smallest number of changes to responding values. The donor imputation was carried out at household level, that is, a nearest neighbour household was identified in the same geographic area as the recipient household, which

was selected to impute all missing or invalid variables under the minimum change principle for all people within a household.

This chapter will develop a nearest neighbour imputation approach, using information from any respondents in a partially responding household to define distance measures.

Section 6.2 describes existing and proposed donor imputation strategies. Section 6.3 describes the simulation study, where the nonresponse mechanisms and outcome variables are unchanged from Chapters 3, 4 and 5 to allow comparison with the donor methods of this chapter. Section 6.4 concludes by contrasting the performance of the proposed donor methods with the findings from Chapters 3, 4 and 5, to compare the benefits of household-based donor imputation approaches to the use of linear mixed models and generalized linear mixed models.

6.2 Donor imputation methods

There are a wide variety of existing donor imputation methods. Two fundamental methods which are routinely used, random donor and class donor, are described in Section 2.4.3. These methods will be used a point of comparison for the proposed household data imputation methods, using only the information about the nonrespondent and ignoring the household structure.

6.2.1 Proposed imputation methods

Nearest Significant Other (notated NSO)

The nearest neighbour method described in Section 2.4.3 can be adapted for household survey data by considering the auxiliary variable, x_{ij}, to be the value of y for a responding person in the household, ie $x_{ij} = y_{i'j}$. The responding person in the context of a two-person household will be called the 'significant other' and therefore the method can be thought of as looking for nearest significant others as the criteria for searching for a donor for the nonrespondent. The nearest neighbour (donor) household, l is the fully responding household containing a person with the closest value of y_{kl} to the respondent $y_{i'j}$ in the recipient household. The variable of interest for the other person in this household, $y_{k'l}$, is then imputed for the nonrespondent. This method, when using a variable which is not strictly continuous, may result in multiple closest donors, from which a random donor would be selected.

This approach can also be applied to binary variable imputation. There would usually be a large set of equally nearest neighbours for binary variables which all have the required binary value and therefore are equidistant from the co-householder. Rather than just selecting a random donor from this pool, one or more additional covariates, such as age and sex, can be used

in addition to the response variable of the respondent, to select a nearest neighbour. This requires the use of a multivariate distance measure, such as the Mahalanobis distance as described in Section 2.4.3. Let $\mathbf{x}_{ij}$ be the covariate vector for person i in household j, which includes the response variable for the significant other in the household, $y_{i'j}$ as one of the x's, and attributes of the respondent, and potentially nonrespondent as other x variables. The distance function to minimise to select the nearest neighbour is therefore: $(\mathbf{x}_{ij} - \mathbf{x}_{kl})^T \hat{\mathbf{V}}_{ij}^{-1} (\mathbf{x}_{ij} - \mathbf{x}_{kl})$ where $\hat{\mathbf{V}}_{ij}$ is the estimated variance-covariance matrix of $\mathbf{x}_{ij}$. The donor value for y_{ij} is then the y value associated with the respondent for which the distance function for the covariate vector is minimised.

In households containing three or more people, where there is one nonrespondent, the set of response values for the household can be used to define the covariate vector used to find the nearest neighbour. For a non-responding person ij the covariate vector $\mathbf{x}_{ij}$ is $(y_{1j}, ..., y_{n_{rj},j})$ where n_{rj} is the number of respondents in the household. A multivariate measure can be used in an analagous way to above to calculate the distance measure between sets of respondents in different households of the same size. This covariate vector can also be extended to incorporate information about the non-responding person. For more than one nonrespondent in a household, a donor household can still be identified using this method by defining a single covariate vector

for the set of nonrespondents consisting of auxiliary information about the nonrespondents and the variable of interest for respondents (and potentially other variables) in the household in a single vector.

This method will be evaluated in the case of two person households where there is either one or zero nonrespondent.

Nearest Significant Other with residual (notated NSO resid)

Another way to incorporate additional covariates is as follows. A linear model $y = \boldsymbol{\beta}^T \mathbf{x}$ results in residuals $r_{ij} = y_{ij} - \hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}$. For a nonresponding person ij the nearest neighbour to their significant other (the responding person in their household), $i'j$, would be identified as the closest residual to $r_{i'j}$, say person k' in a fully responding household l. The other person (k) in this household (l) has their residual used when imputing for y_{ij} over its predicted value: $\hat{y}_{ij} = r_{kl} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}$.

Household respondent (notated hh resp) This method is deterministic and results in a single impute, $\hat{y}_{ij} = y_{i'j}$, the significant other (respondent) in the household. While this method will impute unrealistically similar people within households, it will provide a useful point of comparison to the other donor methods.

6.2.2 Application of donor imputation methods in simulation study

The covariates chosen for the nearest significant other impute were age and sex, of both respondent and nonrespondent, in addition to the response of the co-householder. These were chosen to be consistent with the linear and generalized linear models in earlier chapters to provide a fair comparison.

The imputed values will hopefully come from the distribution of $Y_{2j}|x_{1j}, x_{2j}, y_{1j}$, which will give more realistic within-household dependencies, particularly when there are complex relationships between $(Y_{1j}, \mathbf{x}_{1j})$ and $(Y_{2j}, \mathbf{x}_{2j})$. For example, if very high income earners tend to live with low earners, while moderate income earners are associated with similar earners (and the ICC isn't high because of the reverse correlations on the extremes) this method should do very well, because the low income associated with the partner of a high income earner will be imputed for a nonresponding low hourly wage rate earner. But a nonresponding middle hourly wage rate will be imputed with a similar hourly wage rate.

The nearest neighbour with co-householder residual for imputing hourly wage, with age and sex as independent variables, can be conceptualised as follows; If the responding person in a household with a nonrespondent earns above average hourly wage for their age group and sex, another household will

be identified with both persons responding, where one of those people earn above the hourly wage rate of their age by sex peer group by approximately the same amount. The residual for the other person from this household will be used, for example they may also earn above their peer hourly wage rate, and this residual will be applied to the nonrespondent to vary their model prediction.

6.3 Results

The results below are divided into four parts. Firstly the proposed donor imputation strategies for household survey data (other respondent in household, nearest significant other, and nearest significant other residual) were compared to standard donor imputation strategies (random donor and random within-class donor) for continuous variables. The evaluation criteria of Chapter 3 are used, that is RRMSE and relative bias of the imputed values, resulting ICC, and bias of estimated mean and variance. The same imputation methods are then compared for binary data, using evaluation criteria RRMSE and relative bias of imputed values, and percentage responding the same within households. The second last section relates the household donor strategies back to the results from Chapters 3 and 4 to assess the advantages and disadvantages of household donor methods compared to deterministic

and stochastic linear and linear mixed models for continuous variables. Finally, household donor strategies are compared to stochastic generalized linear mixed model imputed values for the binary variables Y_1 - voting intention labour and Y_2 - employed.

6.3.1 Nearest Significant Other compared to other donor imputed values for continuous variables

Table 6.1 shows the predictive accuracy measured by RRMSE for imputing the other person in the household, and the two nearest neighbour household imputed values, compared to random donor methods and within-class donors using age by sex to define classes. Imputing the respondent from the household (hh resp) results in the lowest RRMSEs, the accuracy results best when the ICC is highest, when people within households are most similar. Aside from this simplistic impute, the nearest neighbour household methods are more accurate than the donor and class donor imputed values, particulary when there is stronger clustering within households, i.e. $\rho = 50\%$ or 85%. The nearest significant other method results in the best RRMSE overall. This method resulted in similar RRMSEs to the nearest significant other method based on the model residuals when $\rho = 20\%$ or 50%, but was more accurate when $\rho = 85\%$.

Table 6.1: RRMSE (%) - imputation using nearest significant other methods compared to other donor methods

NR model	ρ(%)	random donor	class donor	hh resp	NSO	NSO resid
hh MCAR pers MCAR	20	79.1	77.8	72.7	78.0	77.7
	50	82.6	76.7	57.1	69.1	69.6
	85	85.4	84.1	32.1	45.2	49.0
hh MCAR pers MAR	20	81.3	77.8	70.0	79.2	78.7
	50	85.1	76.6	59.4	70.9	71.7
	85	88.1	88.1	33.7	47.8	52.4
hh MCAR pers MNAR	20	81.5	79.9	77.3	78.5	77.5
	50	83.5	78.8	61.2	69.7	70.2
	85	85.1	84.2	34.7	48.0	51.6
hh MNAR pers MCAR	20	80.2	78.6	78.8	76.7	75.2
	50	82.7	77.7	61.9	68.4	68.4
	85	85.2	84.0	34.9	47.6	51.2
hh MNAR pers MNAR	20	80.9	79.5	80.3	76.8	75.2
	50	80.9	76.6	63.0	66.9	66.5
	85	83.1	81.9	35.7	48.2	50.9

Predictive accuracy is also assessed by the relative bias of imputed values under household and standard donor imputation methods, shown in Table 6.2. The donor and class donor methods have low relative bias under the MCAR and MAR nonresponse mechanisms, ranging from -1.4% to 5.1% but the bias increases under the MNAR scenarios, with relative bias between -4.5% and -12.0%. All three household imputation methods are an improvement over donor and class donor imputed values when nonresponse is entirely MCAR, with an overall lower level of bias. When households are MCAR but persons are MAR, the household imputation methods are similar in bias to the standard donor methods. Under the MNAR scenarios the

Table 6.2: Relative Bias (%) - imputation using nearest significant other methods compared to other donor methods

NR model	ρ(%)	random donor	class donor	hh resp	NSO	NSO resid
hh MCAR pers MCAR	20	-1.4	0.4	0.2	0.0	0.4
	50	1.3	0.3	0.1	0.2	0.0
	85	2.0	2.4	0.1	-0.1	-0.4
hh MCAR pers MAR	20	1.5	1.2	-4.5	3.2	-1.1
	50	5.1	0.7	1.1	3.1	-0.1
	85	3.9	1.8	-0.2	-0.3	-1.4
hh MCAR pers MNAR	20	-8.6	-6.8	-4.4	-8.3	-7.7
	50	-6.7	-6.7	-2.8	-7.1	-7.0
	85	-6.2	-5.3	-0.9	-4.1	-4.5
hh MNAR pers MCAR	20	-5.6	-3.7	0.3	-5.4	-5.0
	50	-4.9	-5.0	0.1	-5.1	-5.2
	85	-5.5	-4.5	0.1	-3.4	-3.7
hh MNAR pers MNAR	20	-12.0	-10.1	-4.2	-13.0	-12.3
	50	-10.8	-10.4	-2.0	-11.4	-11.1
	85	-10.9	-9.7	-0.2	-7.3	-7.6

other household respondent impute is excellent for relative bias, less than 5% across all scenarios and ICC levels. The other two household donor imputed values give a small gain in bias under the highest ρ but are no improvement under low or moderate ρ.

The relative bias of the mean estimate using the various donor methods is tabulated in Appendix D, Table 7.5. It reflects the findings in relation to the relative bias of the imputed values, that is the household respondent donor is the superior imputation method, and the other household donor imputation methods are an improvement under the highest ICC, but not much different to the donor and class donor methods under low and moderate ρ. In all cases

the relative bias of the mean estimate was less than 5%.

Donor methods are now compared to the stochastic BLUP imputed values for retaining within-household clustering, shown in Table 6.3.

Table 6.3: Expected value of estimated ICC - imputation using nearest significant other methods compared to other donor methods

NR model	ρ(%)	random donor	class donor	hh resp	NSO	NSO resid
hh MCAR pers MCAR	20	10.1	10.4	59.6	19.3	18.7
	50	24.2	27.7	75.0	50.5	49.0
	85	40.6	43.1	92.5	83.7	80.6
hh MCAR pers MAR	20	11.4	13.2	51.8	18.6	20.4
	50	26.5	31.2	75.5	53.8	52.8
	85	42.0	44.3	92.5	84.0	81.1
hh MCAR pers MNAR	20	9.3	9.7	66.7	18.5	18.0
	50	18.7	22.4	80.9	46.5	44.9
	85	29.0	32.1	94.8	78.8	75.7
hh MNAR pers MCAR	20	8.3	9.1	69.7	17.4	17.0
	50	18.0	21.5	81.9	46.2	44.6
	85	28.2	31.7	94.9	78.6	75.5
hh MNAR pers MNAR	20	7.7	8.9	75.0	16.5	16.2
	50	14.6	17.9	86.1	40.1	39.2
	85	21.5	25.5	96.4	70.9	68.2

With only 25% nonresponse, both the donor and class donor imputed values result in a weakening of the ICC by around 50% under low ICC, but by as much as 75% under the highest ICC. This means that the households will not be representing the true clustering present in the sample.

The limitations of the household respondent impute are clear under this criteria, resulting in imputed values which are too similar to the respondent in the household, and resulting in households which are too similar, with

ICCs too high in all scenarios.

The nearest significant other and residual methods are excellent for reproducing clustering. In particular, the nearest significant other method results in the most accurate estimates of ICC. This method estimates the ICC to within 3.5% under all levels of ρ with MCAR nonresponse, and within 7.6% when nonresponse is MAR. The nearest significant other method is also quite reasonable for MNAR scenarios. When nonresponse for households is MCAR and persons MNAR ICCs are within 7.5% of the true values, within approximately 15% when households are MNAR and persons MCAR, and within 20% when both households and persons are MNAR. These are substantial gains over standard donor imputation methods.

Lastly, the variance performance is assessed in Table 6.4 which measures estimation accuracy in terms of the relative bias of the variance estimate using both household and standard donor methods.

Under MNAR scenarios the variance is almost always underestimated, which is expected as the MNAR mechanism was specified in such a way that individuals with high incomes are less likely to respond.

For example under the nearest neighbour household residual method, if a high income earner is a nonrespondent, their residual is not available in the pool of donor residuals. Having a pool of donor residuals which excludes a large number of high residuals will result in a smaller distribution of residuals

Table 6.4: Relative Bias (%) of Estimated Variance - imputation using nearest significant other methods compared to other donor methods

NR model	ρ(%)	random donor	class donor	hh resp	NSO	NSO resid
hh MCAR pers MCAR	20	-1.8	-0.9	0.2	-0.4	0.9
	50	2.4	-1.5	0.1	-0.3	1.2
	85	3.5	4.9	0.1	-1.4	0.7
hh MCAR pers MAR	20	-0.1	-0.4	-4.5	-0.6	1.6
	50	3.6	-1.8	-0.7	0.9	4.2
	85	-0.9	2.4	-2.5	-4.4	-1.1
hh MCAR pers MNAR	20	-16.9	-16.3	-10.7	-18.3	-17.7
	50	-13.4	-15.9	-7.8	-17.7	-16.4
	85	-11.4	-9.6	-7.4	-14.4	-12.6
hh MNAR pers MCAR	20	-9.2	-8.5	-2.3	-11.4	-11.2
	50	-8.3	-11.0	0.6	-12.9	-12.2
	85	-9.3	-7.6	0.0	-12.5	-10.6
hh MNAR pers MNAR	20	-21.5	-20.9	-9.4	-26.0	-25.8
	50	-15.5	-17.8	-0.6	-23.4	-22.3
	85	-13.7	-12.0	2.4	-21.2	-19.6

overall, and hence a lower variance in the resulting imputed values using these residuals. This is likely the cause for example of the 18.3% and 26.0% underestimates of variance when $p = 20\%$, for the two scenarios with persons MNAR.

6.3.2 Nearest Significant Other compared to donor imputed values for binary variables

This section includes the results from comparing existing donor imputation methods with proposed donor imputation methods making use of household

characteristics. The existing donor methods are a random donor and random class donor. The proposed method uses the multivariate Mahalanobis distance measure to calculate a nearest neighbour based on the age and sex of both household members and the outcome variable for the responding person in the household. This is also contrasted with the multivariate Mahalanobis distance based solely on the known characteristics of the nonrespondent, in this instance their age and sex. These methods are evaluated on a reduced set of criteria (RRMSE, Relative Bias and percentage of households with the same response) for both binary variables, labour and employed.

Firstly, to assess predictive accuracy the RRMSE and Relative Bias are shown, for each the standard donor imputation methods (random donor and random class donor), then for the nearest neighbour and nearest significant other methods.

Tables 6.5 shows the RRMSE for the response variables Y_1 (labour) and Y_2 (employed). The nearest neighbour person donor impute for Y_1 (whether vote labour) is only slightly better for RRMSE than the random donor or class donor methods. However, when this method is extended to incorporate the household attributes, where age group and sex of both household members are matching, there is an improvement in accuracy. RRMSE is reduced by between 13 and 16% across all nonresponse mechanisms. The accuracy for imputed values of Y_2 (employed) had slightly different results. The age group

Table 6.5: Relative Root Mean Square Error (%) - donor imputation methods for Y_1 - voting intention labour and Y_2 - employed

hh	pers	random donor	class donor	NN	NSO
vote labour					
MCAR	MCAR	234·5	245·0	242·3	208·0
MCAR	MAR	231·4	233·1	228·0	196·4
MCAR	MNAR	223·0	220·3	219·8	189·8
MNAR	MCAR	227·7	226·9	225·3	194·3
MNAR	MNAR	208·1	207·6	206·3	178·5
employed					
MCAR	MCAR	112·2	94·2	89·1	85·5
MCAR	MAR	106·6	87·5	83·0	79·5
MCAR	MNAR	105·5	88·3	84·4	80·7
MNAR	MCAR	107·4	90·2	85·9	82·3
MNAR	MNAR	101·6	84·9	81·8	78·4

by sex class donor resulted in more accurate imputed values than the random donor by $16-18\%$. This may mean that the variables age group and sex are more useful predictors of whether employed than of voting intention being labour. As with voting intention, both of the nearest neighbour imputed values resulted in further accuracy improvements.

Table 6.6 shows the relative bias of the imputed values for the outcome variables voting intention and employed under the various donor imputation methods.

For voting intention of labour both the random donor and class donor have similar levels of relative bias in the imputed values, with poorer results under the MNAR scenarios, particularly when persons are MNAR. The near-

Table 6.6: Relative Bias (%) - donor imputation methods for Y_1 - voting intention labour and Y_2 - employed

hh	pers	random donor	class donor	NN	NSO
vote labour					
MCAR	MCAR	1·9	3·2	−8·7	−2·5
MCAR	MAR	−6·1	2·2	−9·3	−3·1
MCAR	MNAR	−11·4	−10·6	−23·3	−12·8
MNAR	MCAR	−7·9	−8·6	−20·7	−9·5
MNAR	MNAR	−20·6	−21·5	−30·0	−19·7
employed					
hh	pers	random donor	class donor	NN	NSO
MCAR	MCAR	−0·2	0·0	−6·9	−0·8
MCAR	MAR	−6·3	0·2	−5·9	−0·6
MCAR	MNAR	−7·3	−5·3	−10·9	−5·6
MNAR	MCAR	−5·1	−3·4	−9·5	−4·1
MNAR	MNAR	−11·7	−7·9	−13·6	−8·5

est neighbour person donor impute has even poorer bias, while the nearest neighbour household impute is similar to the random and class donor imputed values. For the variable employed, there is again an improvement in accuracy for the class donor method over the random donor. The bias issues still exist for the nearest neighbour donor methods, though the size of the bias isn't as high as for voting intention.

The clustering is assessed in Table 6.7 by looking at the percentage of households with the same voting intention or employed status.

The nearest neighbour person donor imputed values resulted in lower clustering than what was actually present within households. The estimates were poorer for the nearest neighbour person impute than either donor or

Table 6.7: Households with same voting intention of labour (%) or same employment status (%)- donor imputation methods

hh	pers	actual	random donor	class donor	NN	NSO
vote labour						
MCAR	MCAR	79·3	70·9	70·7	63·6	80·0
MCAR	MAR	79·3	71·2	71·0	67·8	80·0
MCAR	MNAR	79·3	70·8	71·1	62·9	83·8
MNAR	MCAR	79·3	70·6	70·6	63·4	80·0
MNAR	MNAR	79·3	71·0	71·1	62·9	80·0
employed						
MCAR	MCAR	72·2	62·4	67·3	63·6	72·0
MCAR	MAR	72·2	62·4	67·3	67·8	72·0
MCAR	MNAR	72·2	62·3	67·2	62·9	71·9
MNAR	MCAR	72·2	62·3	67·2	63·4	71·9
MNAR	MNAR	72·2	62·3	67·2	62·9	71·7

class donor methods. This was corrected by use of the nearest neighbour household donor imputed values, where the clustering after imputation is very close to the true values.

Clustering for employment status is also reproduced well by the nearest neighbour household donor imputed values, and poorly by the other methods, which underestimate the proportion of households with the same employment status.

The donor imputation methods for binary variables were then assessed for estimation accuracy, with the results for the relative bias of the population proportion shown in Table 7.6 of Appendix D. The results reflected the evaluation of the relative bias for individual imputed values. The bias in the

proportions is poorer for the nearest neighbour person donor imputed values but this is corrected when the household attributes are included in defining the nearest neighbour.

6.3.3 Comparison of donor methods with linear methods for continuous variables

The stochastic BLUPs are now contrasted to donor imputation methods.

Predictive accuracy - RRMSE and Relative Bias

Table 6.8 shows the predictive accuracy measured by RRMSE for the stochastic single-level and multilevel BLUP imputed values compared to donor methods for imputing hourly wage rate. Age by sex are used as covariates or classes in all models.

Comparing the best of the donor methods (nearest significant other) with the better of the linear methods (stochastic ML BLUP), the accuracy of both sets of imputed values is of a similar level across all scenarios. For the MCAR and MAR scenarios with $\rho = 85\%$ the nearest significant other method is slightly better than the ML BLUP, but for the MNAR mechanisms the ML BLUP is slightly more accurate than the nearest significant other method.

Predictive accuracy is also assessed by comparing the relative bias of imputed values under donor methods and BLUP imputation methods, shown

Table 6.8: Predictive accuracy (RRMSE %) for imputing *hourly wage rate* using linear models compared to donor methods

NR model		ρ(%)	class donor	NSO	NSO resid	SL BLUP	ML BLUP
hh	MCAR	20	77.8	78.0	77.7	78.9	77.7
pers	MCAR	50	76.7	69.1	69.6	78.2	70.3
		85	84.1	45.2	49.0	80.3	50.7
hh	MCAR	20	77.8	79.2	78.7	80.2	81.4
pers	MAR	50	76.6	70.9	71.7	79.5	72.1
		85	88.1	47.8	52.4	83.2	52.2
hh	MCAR	20	79.9	78.5	77.5	80.5	77.3
pers	MNAR	50	78.8	69.7	70.2	79.6	68.6
		85	84.2	48.0	51.6	81.4	47.7
hh	MNAR	20	78.6	76.7	75.2	79.6	74.9
pers	MCAR	50	77.7	68.4	68.4	78.7	67.0
		85	84.0	47.6	51.2	80.8	47.0
hh	MNAR	20	79.5	76.8	75.7	79.9	74.6
pers	MNAR	50	76.6	66.9	66.5	77.6	64.4
		85	81.9	48.2	50.9	79.0	43.7

in Table 6.9.

When comparing the relative bias of imputed values under the nearest significant other method to the ML BLUP, the ML BLUP is slightly better, particularly under the MNAR scenarios where the linear mixed model better accounts for missing higher hourly wage rates than the donor methods. The biggest improvement over donor methods is seen when the ICC is highest and nonresponse is MNAR for both persons and households. Here there are bias issues for donor methods and the SL BLUP. The relative bias ranges between -7.3% and -11.3% when the ICC is 85% for the other imputation methods. In contrast, the bias is only -3.1% for the ML BLUP.

Table 6.9: Predictive accuracy (Relative Bias %) for imputing *hourly wage rate* using linear models compared to donor methods

NR model		ρ(%)	class donor	NSO	NSO resid	SL BLUP	ML BLUP
hh MCAR pers MCAR		20	0.4	0.0	0.4	0.5	0.3
		50	0.3	0.2	0.0	0.4	0.1
		85	2.4	-0.1	-0.4	0.3	0.0
hh MCAR pers MAR		20	1.2	3.2	-1.1	-0.1	0.1
		50	0.7	3.1	-0.1	0.6	0.0
		85	1.8	-0.3	-1.4	0.1	-1.2
hh MCAR pers MNAR		20	-6.8	-8.3	-7.7	-7.1	-6.5
		50	-6.7	-7.1	-7.0	-7.2	-5.3
		85	-5.3	-4.1	-4.5	-7.3	-2.6
hh MNAR pers MCAR		20	3.7	-5.4	-5.0	-3.8	-3.0
		50	-5.0	-5.1	-5.2	-5.1	-3.0
		85	-4.5	-3.4	-3.7	-6.5	-1.6
hh MNAR pers MNAR		20	-10.1	-13.0	-12.3	-10.3	-9.1
		50	-10.4	-11.4	-11.1	-10.4	-7.1
		85	-9.7	-7.3	-7.6	-11.3	-3.1

Estimation accuracy - Relative Bias of mean

Table 7.7 in Appendix B contains shows the relative bias of the mean estimate using donor and linear imputation methods. The bias for estimating the mean was less than 4% across all nonresponse mechanisms and levels of clustering. Overall the ML BLUP is slightly better than both the donor methods and the SL BLUP for estimation of the mean, particularly when person nonresponse is generated MNAR, however there is not a high level of bias to be concerned with.

Estimated intra-class correlation

Donor methods are now compared to the BLUP imputed values for re-

taining within-household clustering, shown in Table 6.10.

Table 6.10: Estimation accuracy for imputing *hourly wage rate* - Estimated intra-class correlation using linear methods compared to donor methods

NR model	ρ(%)	class donor	NSO	NSO resid	SL BLUP	ML BLUP
hh MCAR pers MCAR	20	10.4	19.3	18.7	10.4	18.9
	50	27.7	50.5	49.0	27.4	47.3
	85	43.1	83.7	80.6	45.0	79.9
hh MCAR pers MAR	20	13.2	18.6	20.4	12.8	19.2
	50	31.2	53.8	52.8	30.3	51.0
	85	44.3	84.0	81.1	46.5	81.1
hh MCAR pers MNAR	20	9.7	18.5	18.0	9.5	21.0
	50	22.4	46.5	44.9	22.1	49.6
	85	32.1	78.8	75.7	33.1	83.0
hh MNAR pers MCAR	20	9.1	17.4	17.0	9.0	21.6
	50	21.5	46.2	44.6	21.3	50.3
	85	31.7	78.6	75.5	32.8	83.3
hh MNAR pers MNAR	20	8.9	16.5	16.2	8.3	24.1
	50	17.9	40.1	39.2	18.2	52.6
	85	25.5	70.9	68.2	26.2	86.3

Under MCAR and MAR nonresponse there is no clear best performer for ICC out of the best donor method (nearest neighbour household) and linear method (ML BLUP). When nonresponse is MNAR the multilevel linear imputed values outperform the nearest neighbour household donor method for reproducing clustering. Both the nearest neighbour household and linear mixed imputation method perform very well relative to the class donor and single-level BLUP, but the ML BLUP more closely retains within-household clustering across all levels of ICC under each MNAR scenario.

Relative Bias of Estimated variance

Finally, the performance in preserving variance is assessed in Table 6.11.

Table 6.11: Estimation accuracy for imputing *hourly wage rate* - Relative Bias (%) of Estimated Variance using linear models compared to donor methods

NR model	ρ(%)	class donor	NSO	NSO resid	SL BLUP	ML BLUP
hh MCAR pers MCAR	20	-0.9	-0.4	0.9	0.2	0.3
	50	-1.5	-0.3	1.2	0.2	1.0
	85	4.9	-1.4	0.7	0.2	2.1
hh MCAR pers MAR	20	-0.4	-0.6	1.6	1.9	4.3
	50	-1.8	0.9	4.2	1.7	3.2
	85	2.4	-4.4	-1.1	-3.4	-0.4
hh MCAR pers MNAR	20	-16.3	-18.3	-17.7	-15.7	-18.0
	50	-15.9	-17.7	-16.4	-15.2	-16.5
	85	-9.6	-14.4	-12.6	-14.0	-8.2
hh MNAR pers MCAR	20	-8.5	-11.4	-11.2	-7.8	-11.5
	50	-11.0	-12.9	-12.2	-10.1	-11.8
	85	-7.6	-12.5	-10.6	-12.3	-6.0
hh MNAR pers MNAR	20	-20.9	-26.0	-25.8	-20.6	-25.7
	50	-17.8	-23.4	-22.3	-16.8	-20.2
	85	-12.0	-21.2	-19.6	-16.5	-8.5

On this criteria the household donor methods and ML BLUP are similar under MCAR and MAR, with not much separating the accuracy of the variance estimates. When nonresponse is MNAR, the ML BLUP is the superior method, but only under the highest levels of clustering.

In summary, across the evaluation criteria, the nearest neighbour donor method and ML BLUP are the best of the donor and linear imputation methods, with the ML BLUP slightly better under MNAR, and the nearest neighbour donor slightly better under MCAR and MAR nonresponse.

Percentage of people on or below Federal Minimum Wage for linear and donor methods

The application of imputation to the percentage of people on or below the FMW was re-visited, with the results from linear and donor methods shown in Tables 6.12 and 6.13.

Table 6.12: Percentage of adults on or below FMW - linear methods

			stochastic		stochastic, log		determistic	
NR model	ρ(%)	Actual	SL BLUP	ML BLUP	SL BLUP	ML BLUP	SL BLUP	ML BLUP
hh MCAR pers MCAR	20	7.7	11.0	11.0	8.3	7.9	5.8	5.8
	50	6.8	10.2	10.1	7.5	7.1	5.1	5.6
	85	6.7	10.2	9.7	7.4	6.8	5.0	7.0
hh MCAR pers MAR	20	7.7	11.0	11.2	8.4	8.1	5.8	5.8
	50	6.8	10.2	10.1	7.5	7.2	5.0	5.9
	85	6.7	9.7	9.4	7.3	6.9	4.8	7.4
hh MCAR pers MNAR	20	7.7	11.0	10.5	8.6	8.1	6.0	6.0
	50	6.8	10.2	9.4	7.8	7.3	5.3	5.7
	85	6.7	10.3	9.0	7.7	6.9	5.1	6.9
hh MNAR pers MCAR	20	7.7	11.1	10.4	8.4	7.9	5.9	5.9
	50	6.8	10.2	9.3	7.7	7.2	5.2	5.6
	85	6.7	10.2	8.9	7.6	6.9	5.1	6.9
hh MNAR pers MNAR	20	7.7	11.0	9.9	8.7	8.1	6.1	6.1
	50	6.8	10.3	8.8	7.9	7.2	5.4	5.7
	85	6.7	10.4	8.4	7.8	6.9	5.3	6.8

The household respondent was the best imputation method for reproducing the percentage on or below FMW, followed by the nearest neighbour household method and stochastic ML BLUP under a log-transform. The household respondent reproduced the percentage on or below FMW with at

Table 6.13: Application - Percentage of adults on or below FMW - donor methods

NR model	ρ(%)	Actual	hh resp	donor	class donor	NSO	NSO resid
hh MCAR pers MCAR	20	7.7	7.7	8.2	8.0	8.1	8.5
	50	6.8	6.8	7.4	7.1	7.2	8.2
	85	6.7	6.6	7.3	7.1	6.8	8.2
hh MCAR pers MAR	20	7.7	7.7	8.0	8.2	7.7	9.3
	50	6.8	6.6	7.1	7.2	6.8	8.6
	85	6.7	6.4	6.7	7.1	6.4	8.4
hh MCAR pers MNAR	20	7.7	7.9	8.5	8.3	8.4	8.7
	50	6.8	6.9	7.7	7.4	7.3	8.3
	85	6.7	6.6	7.5	7.3	6.8	8.2
hh MNAR pers MCAR	20	7.7	7.7	8.4	8.1	8.3	8.6
	50	6.8	6.8	7.7	7.4	7.3	8.3
	85	6.7	6.6	7.5	7.3	6.8	8.2
hh MNAR pers MNAR	20	7.7	7.9	8.7	8.4	8.7	8.9
	50	6.8	6.9	7.9	7.5	7.4	8.3
	85	6.7	6.6	7.8	7.6	6.8	8.0

most 2.5% relative bias over all nonresponse models and levels of ICC.

Figure 6.1 shows the estimated percentage on or below FMW for deterministic and stochastic BLUPs and donor methods.

6.3.4 Comparison of donor methods and stochastic generalized linear mixed methods for binary variables

This last subsection of results is a comparison of the evaluation criteria on donor and linear methods for binary variables. The RRMSE, Relative Bias

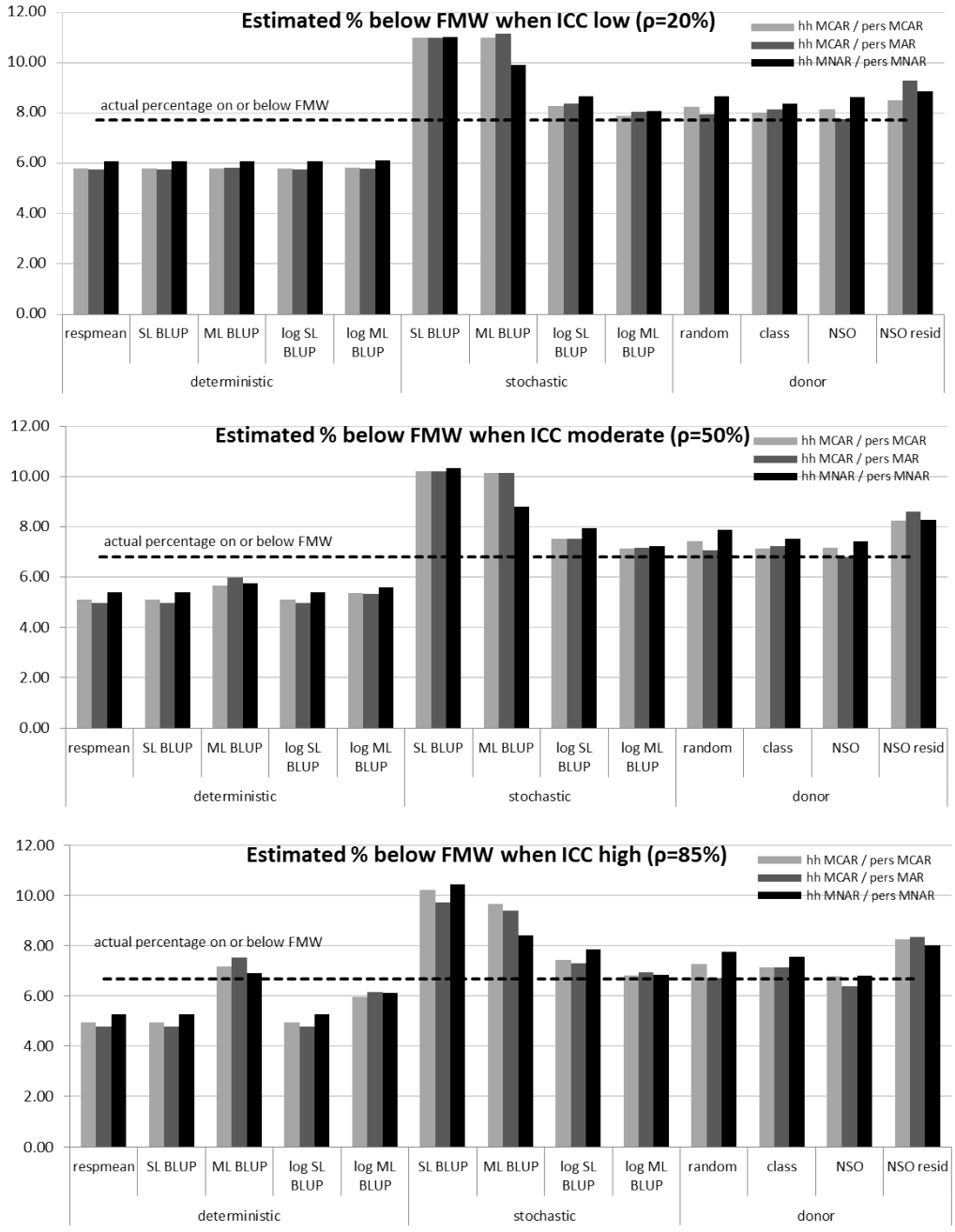

Figure 6.1: Simulation study - estimated percentage on or below Federal Minimum Wage with stochastic and donor imputed values

and proportion of households with the same response are put side-by-side for the household donor methods and linear mixed imputation methods using a single stochastic impute under both a probit and logit model, for both variables voting intention and employment status.

Table 6.14: Relative Root Mean Square Error (%) using linear models compared to donor methods for Y_1 - labour and Y_2 - employed

NR model		donor		logit BLUPs		probit BLUPs	
hh	pers	NN	NSO	SL	ML	SL	ML
vote labour							
MCAR	MCAR	242·3	208·0	242·2	218·0	242·5	217·8
MCAR	MAR	228·0	196·4	232·6	206·7	229·9	207·4
MCAR	MNAR	219·8	189·8	219·0	198·8	220·0	199·5
MNAR	MCAR	225·3	194·3	224·5	202·6	223·8	203·6
MNAR	MNAR	206·3	178·5	205·9	186·4	203·4	188·4
employed							
MCAR	MCAR	89·1	85·5	103·0	101·1	104·4	101·4
MCAR	MAR	83·0	79·5	96·2	94·4	96·7	94·9
MCAR	MNAR	84·4	80·7	97·1	95·4	98·5	95·7
MNAR	MCAR	85·9	82·3	98·8	97·2	100·3	97·5
MNAR	MNAR	81·8	78·4	93·4	92·3	95·6	92·6

Table 6.14 shows that the nearest neighbour household donor method results in the most accurate imputed values, an improvement over either the logit or probit multilevel imputation methods. This holds for both response variables and across all nonresponse mechanisms.

Table 6.15 shows the relative bias of the imputed values for donor imputed values compared to the GLM and GLMM imputed values. The nearest neighbour household donor and multilevel logit model are similar for relative bias

Table 6.15: Relative Bias of imputed values (%) using linear models compared to donor methods for Y_1 - labour and Y_2 - employed

NR model		donor		logit BLUPs		probit BLUPs	
hh	pers	NN	NSO	SL	ML	SL	ML
vote labour							
MCAR	MCAR	−8·7	−2·5	0·9	0·5	1·4	2·6
MCAR	MAR	−9·3	−3·1	2·2	−0·8	−4·8	−0·6
MCAR	MNAR	−23·3	−12·8	−12·9	−11·9	−15·3	−9·9
MNAR	MCAR	−20·7	−9·5	−9·7	−6·7	−14·0	−7·6
MNAR	MNAR	−30·0	−19·7	−20·3	−19·2	−29·4	−18·9
employed							
MCAR	MCAR	−6·9	−0·8	0·0	−0·3	−1·2	−0·6
MCAR	MAR	−5·9	−0·6	−0·4	−0·5	−0·3	−1·2
MCAR	MNAR	−10·9	−5·6	−6·2	−7·2	−8·8	−7·6
MNAR	MCAR	−9·5	−4·1	−4·2	−5·2	−7·3	−5·5
MNAR	MNAR	−13·6	−8·5	−9·6	−11·3	−14·4	−11·7

when assessing voting intentions, but the donor method is slightly better for employment status.

The relative bias of the estimates of proportion are shown in Table 6.16. The nearest neighbour household donor method results in similar relative bias to the ML logit or probit methods.

Lastly, the proportion of households with the same value in each household are compared for each of the variables voting intention and employment status across the donor and generalized linear mixed models approaches in Table 6.17.

The true percentage of households with the same voting intention of labour is 79.3%, and this is underestimated when using the nearest neigh-

Table 6.16: Relative Bias of proportion (%) for Y_1 - vote labour and Y_2 - employed using generalized linear and donor imputation methods

NR model		donor		logit BLUPs		probit BLUPs	
hh	pers	NN	NSO	SL	ML	SL	ML
vote labour							
MCAR	MCAR	−2·2	−1·0	−0·2	−0·3	−0·2	0·2
MCAR	MAR	−2·9	−0·9	0·2	−0·6	−1·7	−0·6
MCAR	MNAR	−6·2	−3·9	−3·9	−3·7	−4·8	−3·1
MNAR	MCAR	−5·4	−2·9	−3·1	−2·2	−3·9	−2·4
MNAR	MNAR	−8·7	−6·2	−6·4	−6·1	−9·3	−5·8
employed							
MCAR	MCAR	−1·3	−0·2	0·0	−0·1	−0·3	−0·2
MCAR	MAR	−1·6	−0·2	−0·1	−0·1	−0·1	−0·3
MCAR	MNAR	−2·6	−1·5	−1·6	−1·9	−2·3	−2·0
MNAR	MCAR	−2·2	−1·1	−1·1	−1·3	−1·9	−1·4
MNAR	MNAR	−3·6	−2·3	−2·6	−3·1	−3·9	−3·2

bour person donor imputation method, but is quite well reproduced under the nearest neighbour household imputed values. While the nearest neighbour household method results in a slight over-clustering for MAR, it is generally close to the true proportion of matches across all nonresponse models, and represents clustering much better than the nearest neighbour person impute which is as much as 16.4 percentage points from the true percentage of households with the same voting intention. Under a logit and probit model the ML imputed values are marginally better than the nearest neighbour household imputed values. However, there is not much to separate these three alternative imputation methods.

The proportion of household with the same employment status was also

Table 6.17: Households with same voting intention of labour (%) and employment status (%) - using generalized linear and donor imputation methods

NR model			donor		logit BLUPs		probit BLUPs	
hh	pers	actual	NN	NSO	SL	ML	SL	ML
vote labour								
MCAR	MCAR	79·3	63·6	80·0	71·0	79·0	71·0	79·1
MCAR	MAR	79·3	67·8	83·8	71·0	80·3	71·4	80·1
MCAR	MNAR	79·3	62·9	80·1	71·1	79·1	71·4	79·0
MNAR	MCAR	79·3	63·4	80·0	71·2	79·2	71·2	78·9
MNAR	MNAR	79·3	62·9	80·0	71·3	79·1	71·8	79·0
employed								
MCAR	MCAR	72·2	63·6	72·0	66·2	70·3	65·7	70·5
MCAR	MAR	72·2	67·8	72·0	66·7	70·5	66·6	70·9
MCAR	MNAR	72·2	62·9	71·9	65·8	69·9	65·3	70·1
MNAR	MCAR	72·2	63·4	71·9	66·0	70·0	65·4	70·1
MNAR	MNAR	72·2	62·9	71·7	65·7	69·5	64·8	69·5

assessed under each of the imputation methods. The true value is 72.2% of households with the same status, and again the nearest neighbour person donor imputed values failed to reproduce this level of clustering, but the nearest neighbour household method was much improved and reproduced the estimate within 0.5 percentage points across each nonresponse mechanism. The SL logit and probit imputed values were consistent with the previous variable in their under-estimation of clustering. While the ML imputed values were an improvement on the SL imputed values, they didn't preserve clustering as well as the nearest neighbour household approach.

6.4 Summary of Chapter 6

This chapter contained a thorough examination of both new and existing donor methods in the household survey setting. These methods were contrasted to the linear and generalized linear model imputation methods discussed in Chapters 3 through to 5. The results of this chapter therefore provide an overall assessment of how well donor and model-based imputed values perform for both continuous and binary variables in the household setting under different assumptions about nonresponse and varying levels of within-household clustering.

The proposed donor methods which made use of household characteristics showed promising results, with some substantial gains over standard donor methods demonstrated.

Imputing the response variable from the responding person in the household resulted in low RRMSE and bias and good preservation of variance for both continuous and binary variables, and performed well for identifying the proportion of people at or under the Federal Minimum Wage. However, this approach was particularly poor at preserving ICC. The major drawback is that it clearly creates unrealistic imputed values, with households with a nonrespondent ending up with a perfect correlation of the response variable, distorting the sample ICC.

Both household donor methods were clearly better than donor and class donor imputed values for continuous variables, particularly when the ICC was large. There were improvements to RRMSE, bias, and preservation of ICC. The nearest significant other impute did better than the residual-based variant on this method across all evaluation criteria. In terms of variance estimation, the donor methods all displayed low levels of bias under the MCAR and MAR scenarios, and only the household respondent method made any improvements to the substantial under-estimation of variance under the MNAR nonresponse mechanisms, where it performed remarkably well.

When binary variables were investigated, the findings were consistent across all criteria and nonresponse mechanisms. The nearest neighbour household imputation method resulted in superior imputed values for binary variables compared to other household donor imputation methods as well as the generalized linear model imputed values under logit or probit models. This method also resulted in the most accurate imputed values, with smaller RRMSE than all other methods for both binary variables. When estimating population proportions, the nearest neighbour household donor method resulted in smaller relative bias than the ML multiple probit imputed values, and was also slightly better than the ML logit multiple imputed values. The results for clustering also pointed to this being the superior method, with similar results to the ML multiple imputed values under the probit or

logit models for one binary variable but superior to them for reproducing the proportion of households with the same value for the other binary response variable.

Chapter 7

Conclusions

7.1 Summary of findings

Household surveys data are an example of hierarchically structured data, characterised by small cluster sizes, and data items with varying levels of intra-household correlation. This thesis explored several methods for imputing missing item-level data by exploiting the household structure in different ways. These included deterministic imputed values based on a SL BLUP and a ML BLUP, a single-level impute incorporating information about a respondent in the household, stochastic BLUPs and multiple imputation. For binary variables GLM and GLMM imputed values were derived, using single or multiple stochastic imputed values. Various donor imputation methods were developed incorporating household information, and assessed against

the linear and generalised linear model imputed values.

While some authors have investigated the use of mixed models in imputation, and some specialist statistical software packages are beginning to offer this capability, there was no strong existing evidence on whether, and when, the additional resources dedicated to the development of these more complex models in household survey data was worthwhile. This thesis therefore considered new and existing imputation methods for their performance under a range of missing data mechanisms (MCAR, MAR, MNAR) and with varying levels of intra-household correlation.

These imputation methods were assessed across a range of criteria relevant to household surveys, including standard accuracy measures such as RRMSE and bias for imputation of the values for individuals, and relative bias for means, but also distributional attributes such as variance estimates and intra-household correlation. Simulation studies using Australian and British data provided examples of the potential gains achievable using each of these methods in a household survey setting.

Deterministic BLUPs

The performance of the ML BLUP compared to a SL BLUP depended on the level of clustering and the nonresponse mechanism. Improvements in accuracy were seen across all nonresponse mechanisms, and reductions in bias

achieved by the ML BLUP over the SL BLUP under information nonresponse mechanisms. The SL BLUP under estimated ICC significantly for all levels of clustering, and this was improved by the ML BLUP, though it tended to impute values that were too similar within households. While gains were made in some areas when using the ML BLUP compared to SL BLUP imputed values, the deterministic imputed values resulted in underestimation of variance. They are therefore not ideal in household surveys, where distributions and relationships between household members are of importance. Stochastic imputation methods were proposed and assessed against the deterministic imputed values to resolve these deficiencies.

Stochastic BLUPs and MI

A stochastic element was introduced to the SL BLUP and ML BLUP imputed values, and both a single and multiple imputed values were derived, under the same nonresponse and clustering scenarios as for the deterministic imputed values. As with the deterministic imputed values, the stochastic ML BLUP resulted in improvements in accuracy and bias which increased with the size of the ICC. Both the stochastic SL BLUP and ML BLUP resulted in improved variance estimates compared to their deterministic versions. However, there was still under-estimation of variance, and this was not generally resolved by the introduction of multiple imputation. With regards to ICC the

stochastic ML BLUP performed very strongly, resulting in accurate representations of within-household clustering across all nonresponse mechanisms and ICC levels, not achieved by the stochastic SL BLUP.

Binary imputed values

Imputation methods for binary variables were developed using both a GLM and a GLMM. Single stochastic imputed values, and multiple imputed values were compared on two binary variables with different levels of within-household clustering. The GLMM resulted in small but consistent accuracy improvements over the GLM imputed values. A further improvement was achieved with multiple rather than single stochastic imputed values. The GLMM resulted in little improvement for the bias of estimates of population proportions, but showed substantial gains for estimating the proportion of households with the same value of the imputation variable within a household. As these estimates were quite accurate there was no additional gain found from the introduction of multiple imputed values.

Donor methods

Several donor methods were proposed making use of household information, imputing the response from another person in the household, a nearest neighbour household imputation method (nearest significant other) and a nearest significant other impute based on the model residual. Imputing the

response from another person in the household led to improvements in accuracy and bias as well as good variance estimates, but resulted in unrealistic imputed values with perfect intra-household correlation distorting the sample ICC. The nearest significant other donor impute was the best of the donor methods, with some good gains over standard donor methods when assessed on accuracy, bias and preservation of ICC. These findings were consistent for continuous and binary variables.

Overall, the multilevel stochastic BLUP, multilevel stochastic BLUP with multiple imputed values, and nearest significant other method result in improvements over existing imputation methods, but when compared to each other were similar, with varying performances across the different evaluation criteria. The improvements over standard methods are greatest when the ICC is large, and when nonresponse is informative. The main benefit compared to non-household approaches is in the preservation of intra-household dependencies, both for binary and continuous variables.

7.2 Further research

In Chapters 3 and 4 imputed values were derived under a log transformation, which is commonly applied when analysing skewed data such as income. A higher level of bias was found in imputed values derived under both a

linear and linear mixed model. The bias corrections applied were able to deal with this issue for single-level imputed values (under non-informative response), but not for multilevel imputed values. Imputed values based on log transformed data should therefore not be used without bias correction, and further investigation would be useful into the cause of, and potential solutions for, bias issues with multilevel imputation using a log transform.

The proposed and existing household imputation methods were assessed in this thesis using a range of nonresponse mechanisms- MCAR, MAR and MNAR, with nonresponse being informative at person-level, or household-level, or both. An alternative would be to model the nonresponse mechanism using a multilevel model, and assess the performance of these imputation methods under further scenarios.

Another area for further research is to build on this work by considering the additional layer of hierarchy present in longitudinal household surveys. In this case an imputation model could build in correlations over time as well as within households. Alternatively, or in addition, geographic regions can be incorporated into a three or more level hierarchy.

Appendix A - Matrix algebra for derivation of BLUP in Chapter 3

Matrix Algebra for BLUPs

Let y_{ij} be the response variable of interest for person i in household j. $\mathbf{y} = (y_{11}, y_{21}, ..., y_{ij}, ..., y_{n_m m})^T$ is the response vector of length n (n=n_o+n_u) and $\mathbf{X}$ be the associated matrix of fixed effects and $\mathbf{Z}$ be the matrix of random effects.

$$
\begin{aligned}
y_{ij[1\text{x}1]} &= \mathbf{X}_{ij[1\text{x}p]}\boldsymbol{\beta}_{[p\text{x}1]} + \mathbf{z}_{ij[1\text{x}q]}\mathbf{u}_{j[q\text{x}1]} + e_{ij[1\text{x}1]} \\
\text{or} \qquad \mathbf{y}_{j[n_j\text{x}1]} &= \mathbf{X}_{j[n_j\text{x}p]}\boldsymbol{\beta}_{[p\text{x}1]} + \mathbf{z}_{j[n_j\text{x}q]}\mathbf{u}_{j[q\text{x}1]} + \mathbf{e}_{j[n_j\text{x}1]} \\
\mathbf{u_j} &\sim N(\mathbf{0}, \mathbf{D}) \\
e_j &\sim N(\mathbf{0}, \boldsymbol{\Omega}_{1,j}) \\
cov(\mathbf{u_j}, \mathbf{e_j}) &= \mathbf{0}
\end{aligned}
$$

In matrix notation:

$$\begin{aligned} \mathbf{Y}_{[\text{nx1}]} &= \mathbf{X}_{[\text{nxp}]}\boldsymbol{\beta}_{[\text{px1}]} + \mathbf{Z}_{[\text{nxqm}]}\mathbf{u}_{[\text{qmx1}]} + \mathbf{e}_{[\text{nx1}]} \\ \mathbf{u} &\sim N(0, \boldsymbol{\Omega}_2) \\ \mathbf{e} &\sim N(0, \boldsymbol{\Omega}_1) \\ cov(\mathbf{u}, \mathbf{e}) &= \mathbf{0} \end{aligned}$$

Where $\mathbf{Z}$ is a block diagonal matrix, with the $\mathbf{z_j}$ matrices on the diagonal, $\mathbf{u}$ is a vector formed by stacking the $\mathbf{u_j}$ vectors vertically.

- $\mathbf{y_o}$ is the observed component of $\mathbf{y}$, an (n_o x 1) vector with associated (n_o x p_1) matrix of covariates $\mathbf{X_o}$.
- $\mathbf{y_u}$ is the unobserved component of $\mathbf{y}$, an (n_u x 1) vector with associated (n_u x p_1) matrix of observed covariates $\mathbf{X_u}$.

The *linear* predictor for $\mathbf{y_u}$ is $\hat{\mathbf{y}}_\mathbf{u} = \mathbf{W}^T\mathbf{y_o}$ where $\mathbf{W}$ is an (n_o x n_u) matrix of weights. The BLUP for $\mathbf{y_u}$ given $\mathbf{X_u}$ is found by calculating $\mathbf{W}$ to minimise the prediction variance $var(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u})$ (*best*) with $E(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u}) = 0$ (*unbiased*).

Let $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $var(\mathbf{y_o}) = \boldsymbol{\Omega_o}$ be the (. x .) covariance matrix for $\mathbf{y_o}$ and $var(\mathbf{y_u}) = \boldsymbol{\Omega_u}$ be the (. x .) covariance matrix for $\mathbf{y_u}$.

As an unbiased predictor $E(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y}_\mathbf{u}) = \mathbf{0}$, therefore:

$$\begin{aligned} E(\mathbf{W^T y_o} - \mathbf{y_u}) &= \mathbf{0} \\ \mathbf{W^T}E(\mathbf{y_o}) - E(\mathbf{y_u}) &= \mathbf{0} \\ \mathbf{W^T X_o}\boldsymbol{\beta} - \mathbf{X_u}\boldsymbol{\beta} &= \mathbf{0} \\ (\mathbf{W^T X_o} - \mathbf{X_u})\boldsymbol{\beta} &= \mathbf{0} \end{aligned}$$

For this to be true for any $\boldsymbol{\beta}$, $(\mathbf{W^T X_o} - \mathbf{X_u}) = \mathbf{0}$.

To find the BLUP for $\mathbf{y_u}$ the prediction variance is minimised:

$$\begin{aligned} var(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u}) &= var(\mathbf{W^T y_o} - \mathbf{y_u}) \\ &= var(\mathbf{W}^T\mathbf{y_o}) + var(\mathbf{y_u}) - 2\mathbf{W}^T cov(\mathbf{y_o}, \mathbf{y_u}) \\ &= \mathbf{W}^T\boldsymbol{\Omega}_\mathbf{o}\mathbf{W} + \boldsymbol{\Omega}_\mathbf{u} - 2\mathbf{W}^T cov(\mathbf{y_o}, \mathbf{y_u}) \\ &= \mathbf{W}^T(\boldsymbol{\Omega}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y_o}, \mathbf{y_u})) + \boldsymbol{\Omega}_\mathbf{u} \end{aligned}$$

To find $\mathbf{W}$ the prediction variance is minimised subject to $(\mathbf{W}^T\mathbf{X_o} -$

$\mathbf{X_u}) = \mathbf{0}$ by the method of Lagrange multipliers (reference required?):

$$\begin{aligned} L(\mathbf{W}, \boldsymbol{\lambda}) &= var(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u}) + (\mathbf{W}^T\mathbf{X_o} - \mathbf{X_u})\boldsymbol{\lambda} \\ &= \mathbf{W}^T(\boldsymbol{\Omega}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y_o}, \mathbf{y_u})) + \boldsymbol{\Omega}_\mathbf{u} + (\mathbf{W}^T\mathbf{X_o} - \mathbf{X_u})\boldsymbol{\lambda} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W}} &= \mathbf{0} \\ 2\boldsymbol{\Omega}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y_o}, \mathbf{y_u}) + \mathbf{X_o}\boldsymbol{\lambda} &= \mathbf{0} \\ 2\boldsymbol{\Omega}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y_o}, \mathbf{y_u}) - 2\mathbf{X_o}\boldsymbol{\lambda}^* &= \mathbf{0} \qquad where \quad \boldsymbol{\lambda}^* = -\frac{\boldsymbol{\lambda}}{2} \\ -\boldsymbol{\Omega}_\mathbf{o}\mathbf{W} + cov(\mathbf{y_o}, \mathbf{y_u}) + \mathbf{X_o}\boldsymbol{\lambda}^* &= \mathbf{0} \\ \boldsymbol{\Omega}_\mathbf{o}\mathbf{W} &= cov(\mathbf{y_o}, \mathbf{y_u}) + \mathbf{X_o}\boldsymbol{\lambda}^* \\ \mathbf{W} &= \boldsymbol{\Omega}_\mathbf{o}{}^{-1}(\mathbf{X_o}\boldsymbol{\lambda}^* + cov(\mathbf{y_o}, \mathbf{y_u})) \end{aligned}$$

Now $\mathbf{X_u} = \mathbf{W^T X_o}$ as $\frac{\partial L}{\partial \boldsymbol{\lambda}} = 0$ and therefore $\mathbf{X_u^T} = \mathbf{X_o^T W}$. Substituting the

above expression for $\mathbf{W}$ will give an expression for $\boldsymbol{\lambda}^*$:

$$\begin{aligned}
\mathbf{X_u^T} &= \mathbf{X_o^T \Omega_o^{-1}}(\mathbf{X_o}\boldsymbol{\lambda}^* + cov(\mathbf{y_o}, \mathbf{y_u})) \\
\mathbf{X_u^T} &= \mathbf{X_o^T \Omega_o^{-1} X_o}\boldsymbol{\lambda}^* + \mathbf{X_o^T \Omega_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u}) \\
(\mathbf{X_o^T \Omega_o^{-1} X_o})\boldsymbol{\lambda}^* &= \mathbf{X_u^T} - \mathbf{X_o^T \Omega_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u}) \\
\boldsymbol{\lambda}^* &= (\mathbf{X_o^T \Omega_o^{-1} X_o})^{-1}(\mathbf{X_u^T} - \mathbf{X_o^T \Omega_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u}))
\end{aligned}$$

The weights $\mathbf{W}$ forming the BLUP for missing $\mathbf{y_u}$ are then given by:

$$\mathbf{W} = \mathbf{\Omega_o^{-1}}\{\mathbf{X_o}(\mathbf{X_o}^T\mathbf{\Omega_o^{-1}X_o})^{-1}(\mathbf{X_u}^T - \mathbf{X_o}^T\mathbf{\Omega_o^{-1}}cov(\mathbf{y_o}, \mathbf{y_u})) + cov(\mathbf{y_o}, \mathbf{y_u})\}$$

and hence the BLUP for $\mathbf{y_u}$ is given by:

$$\begin{aligned}
\hat{\mathbf{y}}_\mathbf{u} &= \mathbf{W^T y_o} \\
&= \{\mathbf{\Omega_o^{-1}}(\mathbf{X_o}(\mathbf{X_o^T \Omega_o^{-1} X_o})^{-1}(\mathbf{X_u^T} - \mathbf{X_o^T \Omega_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u})) + cov(\mathbf{y_o}, \mathbf{y_u}))\}^T \mathbf{y_o} \\
&= \{\mathbf{\Omega_o^{-1} X_o}(\mathbf{X_o^T \Omega_o^{-1} X_o})^{-1}\mathbf{X_u^T} + \mathbf{\Omega_o^{-1}}[\mathbf{I} - \mathbf{X_o}^T(\mathbf{X_o}^T\mathbf{\Omega_o^{-1}X_o})^{-1}\mathbf{X_o}^T\mathbf{\Omega_o^{-1}}]cov(\mathbf{y_o}, \mathbf{y_u})\}^T\mathbf{y_o}
\end{aligned}$$

This text presents the Best Linear Unbiased Predictor (BLUP) for imputing missing data in a single variable of interest under firstly a single-level

model, then a multilevel model.

The Best Linear Unbiased Predictor (BLUP) for missing observations in variable of interest $\mathbf{y}$ will be derived firstly using a single-level model. This model may be applied to missing household-level data, or to missing person-level data (which has the effect of ignoring the household structure).

Let $\mathbf{y} = (y_1, y_2, ..., y_i, ..., y_n)^T$ be a vector of length n (n=n_o+n_u) and $\mathbf{X}$ be the associated (n x p_1) matrix of fully observed covariates.

- $\mathbf{y_o}$ is the observed component of $\mathbf{y}$, an (n_o x 1) vector with associated (n_o x p_1) matrix of covariates $\mathbf{X_o}$.

- $\mathbf{y_u}$ is the unobserved component of $\mathbf{y}$, an (n_u x 1) vector with associated (n_u x p_1) matrix of known covariates $\mathbf{X_u}$.

The *linear* predictor for $\mathbf{y_u}$ is $\hat{\mathbf{y}}_\mathbf{u} = \mathbf{W}^T\mathbf{y_o}$ where $\mathbf{W}$ is an (n_o x n_u) matrix of weights. The BLUP for $\mathbf{y_u}$ given $\mathbf{X_u}$ is found by calculating $\mathbf{W}$ to minimise the prediction variance $var(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u})$ (*best*) with $E(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u}) = 0$ (*unbiased*).

Let $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $var(\mathbf{y_o}) = \mathbf{V_o}$ be a (n_0 x n_0) covariance matrix and $var(\mathbf{y_u}) = \mathbf{V_u}$ be a (n_u x n_u) covariance matrix .

As an unbiased predictor $E(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y}_\mathbf{u}) = \mathbf{0}$, therefore:

$$\begin{aligned} E(\mathbf{W^T y_o} - \mathbf{y_u}) &= \mathbf{0} \\ \mathbf{W^T}E(\mathbf{y_o}) - E(\mathbf{y_u}) &= \mathbf{0} \\ \mathbf{W^T X_o}\boldsymbol{\beta} - \mathbf{X_u}\boldsymbol{\beta} &= \mathbf{0} \\ (\mathbf{W^T X_o} - \mathbf{X_u})\boldsymbol{\beta} &= \mathbf{0} \end{aligned}$$

For this to be true for any $\boldsymbol{\beta}$, $(\mathbf{W^T X_o} - \mathbf{X_u}) = \mathbf{0}$.

To find the BLUP for $\mathbf{y_u}$ the prediction variance is minimised:

$$\begin{aligned} var(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y_u}) &= var(\mathbf{W}^T\mathbf{y_o} - \mathbf{y_u}) \\ &= var(\mathbf{W}^T\mathbf{y_o}) + var(\mathbf{y_u}) - 2\mathbf{W}^T cov(\mathbf{y_o}, \mathbf{y_u}) \\ &= \mathbf{W}^T\mathbf{V_o}\mathbf{W} + \mathbf{V_u} - 2\mathbf{W}^T cov(\mathbf{y_o}, \mathbf{y_u}) \\ &= \mathbf{W}^T(\mathbf{V_o}\mathbf{W} - 2cov(\mathbf{y_o}, \mathbf{y_u})) + \mathbf{V_u} \end{aligned}$$

To find $\mathbf{W}$ the prediction variance is minimised subject to $(\mathbf{W}^T\mathbf{X_o} - \mathbf{X_u}) = \mathbf{0}$

by the method of Lagrange multipliers (reference required?):

$$\begin{aligned} L(\mathbf{W}, \boldsymbol{\lambda}) &= var(\hat{\mathbf{y}}_\mathbf{u} - \mathbf{y}_\mathbf{u}) + (\mathbf{W}^T\mathbf{X}_\mathbf{o} - \mathbf{X}_\mathbf{u})\boldsymbol{\lambda} \\ &= \mathbf{W}^T(\mathbf{V}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y}_\mathbf{o}, \mathbf{y}_\mathbf{u})) + \mathbf{V}_\mathbf{u} + (\mathbf{W}^T\mathbf{X}_\mathbf{o} - \mathbf{X}_\mathbf{u})\boldsymbol{\lambda} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W}} &= \mathbf{0} \\ 2\mathbf{V}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y}_\mathbf{o}, \mathbf{y}_\mathbf{u}) + \mathbf{X}_\mathbf{o}\boldsymbol{\lambda} &= \mathbf{0} \\ 2\mathbf{V}_\mathbf{o}\mathbf{W} - 2cov(\mathbf{y}_\mathbf{o}, \mathbf{y}_\mathbf{u}) - 2\mathbf{X}_\mathbf{o}\boldsymbol{\lambda}^* &= \mathbf{0} \qquad where \quad \boldsymbol{\lambda}^* = -\frac{\boldsymbol{\lambda}}{2} \\ -\mathbf{V}_\mathbf{o}\mathbf{W} + cov(\mathbf{y}_\mathbf{o}, \mathbf{y}_\mathbf{u}) + \mathbf{X}_\mathbf{o}\boldsymbol{\lambda}^* &= \mathbf{0} \\ \mathbf{V}_\mathbf{o}\mathbf{W} &= cov(\mathbf{y}_\mathbf{o}, \mathbf{y}_\mathbf{u}) + \mathbf{X}_\mathbf{o}\boldsymbol{\lambda}^* \\ \mathbf{W} &= \mathbf{V}_\mathbf{o}^{-1}(\mathbf{X}_\mathbf{o}\boldsymbol{\lambda}^* + cov(\mathbf{y}_\mathbf{o}, \mathbf{y}_\mathbf{u})) \end{aligned}$$

Now $\mathbf{X_u} = \mathbf{W^T X_o}$ as $\frac{\partial L}{\partial \boldsymbol{\lambda}} = 0$ and therefore $\mathbf{X_u^T} = \mathbf{X_o^T W}$. Substituting the

above expression for $\mathbf{W}$ will give an expression for $\boldsymbol{\lambda}^*$:

$$\begin{aligned}
\mathbf{X_u^T} &= \mathbf{X_o^T V_o^{-1}}(\mathbf{X_o}\boldsymbol{\lambda}^* + cov(\mathbf{y_o}, \mathbf{y_u})) \\
\mathbf{X_u^T} &= \mathbf{X_o^T V_o^{-1} X_o}\boldsymbol{\lambda}^* + \mathbf{X_o^T V_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u}) \\
(\mathbf{X_o^T V_o^{-1} X_o})\boldsymbol{\lambda}^* &= \mathbf{X_u^T} - \mathbf{X_o^T V_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u}) \\
\boldsymbol{\lambda}^* &= (\mathbf{X_o^T V_o^{-1} X_o})^{-1}(\mathbf{X_u^T} - \mathbf{X_o^T V_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u}))
\end{aligned}$$

The weights $\mathbf{W}$ forming the BLUP for missing $\mathbf{y_u}$ are then given by:

$$\mathbf{W} = \mathbf{V_o^{-1}}\{\mathbf{X_o}(\mathbf{X_o}^T\mathbf{V_o^{-1}X_o})^{-1}(\mathbf{X_u}^T - \mathbf{X_o}^T\mathbf{V_o^{-1}}cov(\mathbf{y_o}, \mathbf{y_u})) + cov(\mathbf{y_o}, \mathbf{y_u})\}$$

and hence the BLUP for $\mathbf{y_u}$ is given by:

$$\begin{aligned}
\hat{\mathbf{y}}_\mathbf{u} &= \mathbf{W^T y_o} \\
&= \{\mathbf{V_o^{-1}}(\mathbf{X_o}(\mathbf{X_o^T V_o^{-1} X_o})^{-1}(\mathbf{X_u^T} - \mathbf{X_o^T V_o^{-1}} cov(\mathbf{y_o}, \mathbf{y_u})) + cov(\mathbf{y_o}, \mathbf{y_u}))\}^T \mathbf{y_o} \\
&= \{\mathbf{V_o^{-1} X_o}(\mathbf{X_o^T V_o^{-1} X_o})^{-1}\mathbf{X_u^T} + \mathbf{V_o^{-1}}[\mathbf{I} - \mathbf{X_o}(\mathbf{X_o}^T\mathbf{V_o^{-1}X_o})^{-1}\mathbf{X_o}^T\mathbf{V_o^{-1}}]cov(\mathbf{y_o}, \mathbf{y_u})\}^T\mathbf{y_o}
\end{aligned}$$

If $\mathbf{V_o} = \sigma^2\mathbf{I}$ then this simplifies to $\mathbf{W} = \mathbf{X_o}(\mathbf{X_o^T X_o})^{-1}\mathbf{X_u^T}$ and hence $\hat{\mathbf{y}}_\mathbf{u} = \mathbf{W^T y_o} = \mathbf{X_u}(\mathbf{X_o^T X_o})^{-1}(\mathbf{X_o^T y_o})$, which is the ordinary least squares

estimator.

Variance Models

The preceding derivation was non-specific about the form of $\mathbf{V_o}$, the variance matrix for the observed response vector $\mathbf{y_o}$, and as the BLUP includes the term $\mathbf{V_o}$, it must be specified before the BLUP weights can be calculated. Below are alternative models for the variance structure.

Alternative variance structures for a person-level single response variable y_i for person i in household j, ignoring the household clustering:

Model	$Var(y_i)$	$C(y_i, y_{i'})$
1	σ^2	0

Under this model there is no within-household correlation for the variable of interest, hence $\mathbf{V_o} = \sigma^2\mathbf{I}$ and $cov(\mathbf{y_o}, \mathbf{y_u}) = \mathbf{0}$.

Hence the BLUP for $\mathbf{y_u}$ simplifies to:

$$\begin{aligned}
\hat{y}_u &= \{\mathbf{V_o^{-1}X_o}(\mathbf{X_o}^T\mathbf{V_o^{-1}X_o})^{-1}\mathbf{X_u}^T + \mathbf{V_o^{-1}}[\mathbf{I} - \mathbf{X_o}(\mathbf{X_o}^T\mathbf{V_o^{-1}X_o})^{-1}\mathbf{X_o}^T\mathbf{V_o^{-1}}] \\
&\quad cov(\mathbf{y_o}, \mathbf{y_u})\}^T\mathbf{y_o} \\
&= \{(\sigma^2\mathbf{I})^{-1}\mathbf{X_o}(\mathbf{X_o}^T(\sigma^2\mathbf{I})^{-1}\mathbf{X_o})^{-1}\mathbf{X_u}^T \\
&\quad +(\sigma^2\mathbf{I})^{-1}[\mathbf{I} - \mathbf{X_o}(\mathbf{X_o}^T(\sigma^2\mathbf{I})^{-1}\mathbf{X_o})^{-1}\mathbf{X_o}^T(\sigma^2\mathbf{I})^{-1}]\mathbf{0}\}^T\mathbf{y_o} \\
&= \{\mathbf{X_o}(\mathbf{X_o}^T\mathbf{X_o})^{-1}\mathbf{X_u}^T\}^T\mathbf{y_o}
\end{aligned}$$

Appendix B - additional tables from Chapter 3

Table 7.1: Estimation accuracy for imputing *Hourly Wage Rate* - Relative Bias (%) of Estimated Mean

			Deterministic BLUPs			
NR model		ρ(%)	Respmean	SL	SL+	ML
hh MCAR pers MCAR		20	0.0	0.1	0.0	0.1
		50	0.1	0.0	0.0	0.0
		85	0.1	0.0	0.0	0.0
hh MCAR pers MAR		20	0.7	0.0	0.2	0.0
		50	0.9	0.2	-0.2	-0.1
		85	0.5	0.0	-0.2	-0.3
hh MCAR pers MNAR		20	-2.0	-1.9	-2.1	-1.7
		50	-2.0	-1.9	-1.7	-1.4
		85	-2.0	-1.9	-0.7	-0.7
hh MNAR pers MCAR		20	-1.1	-1.0	-1.3	-0.8
		50	-1.5	-1.4	-1.2	-0.8
		85	-1.8	-1.8	-0.5	-0.5
hh MNAR pers MNAR		20	-3.0	-2.9	-3.3	-2.5
		50	-3.1	-2.9	-2.6	-2.0
		85	-3.4	-3.2	-0.9	-0.9

Table 7.2: Estimation accuracy for imputing *hourly wage rate* - relative bias (%) of estimated mean using log transform

		BLUP		BLUP log		BLUP log BC	
NR model	ρ(%)	SL	ML	SL	ML	SL	ML
hh MCAR pers MCAR	20	0·1	0·1	−2·3	−3·3	−0·1	1·2
	50	0·0	0·0	−2·3	−2·8	−0·1	1·8
	85	0·0	0·0	−2·4	−1·9	−0·1	3·0
hh MCAR pers MAR	20	0·0	0·0	−2·1	−3·1	0·0	1·5
	50	0·2	−0·1	−2·0	−2·6	0·1	1·9
	85	0·0	−0·3	−2·3	−2·0	−0·3	2·4
hh MCAR pers MNAR	20	−1·9	−1·7	−4·1	−5·0	−2·0	−0·8
	50	−1·9	−1·4	−4·1	−1·3	−2·0	0·0
	85	−1·9	−0·7	−4·2	−2·7	−2·1	2·0
hh MNAR pers MCAR	20	−1·0	−0·8	−3·3	−4·2	−1·2	0·1
	50	−1·4	−0·8	−3·6	−3·7	−1·5	0·6
	85	−1·8	−0·5	−4·0	−2·5	−1·9	2·2
hh MNAR pers MNAR	20	−2·9	−2·5	−5·0	−5·8	−2·9	−1·8
	50	−2·9	−2·0	−5·0	−4·9	−3·0	−0·8
	85	−3·2	−0·9	−5·4	−3·0	−3·3	1·6

Appendix B - Additional tables for Chapter 4

Table 7.3: Rel. Bias (%) of estimated mean - MI compared to single imputation using SL and ML BLUPs

NR model		ρ(%)	Single Stochastic SL	Single Stochastic ML	MI SL	MI ML	MI log SL	MI log ML
hh MCAR pers MCAR		20	0.1	0.1	-0.5	-0.3	1.8	3.1
		50	0.1	0.0	0.0	0.0	2.2	3.7
		85	0.1	-0.0	-0.1	-0.1	2.0	3.7
hh MCAR pers MAR		20	-0.0	0.0	0.8	0.7	3.0	4.8
		50	0.1	-0.0	0.1	0.0	2.5	3.8
		85	0.0	-0.3	0.1	-0.3	2.0	3.2
hh MCAR pers MNAR		20	-1.9	-1.7	-1.9	-1.6	0.4	1.8
		50	-1.9	-1.4	-1.9	-1.3	0.3	2.0
		85	-1.9	-0.7	-1.8	-0.6	0.4	3.1
hh MNAR pers MCAR		20	-1.0	-0.8	-0.8	-0.6	1.3	2.4
		50	-1.3	-0.8	-1.5	-0.9	0.7	2.1
		85	-1.7	-0.4	-1.8	-0.6	0.4	3.1
hh MNAR pers MNAR		20	-2.8	-2.5	-3.1	-2.9	-1.1	-0.4
		50	-2.9	-2.0	-2.9	-2.1	-0.7	1.1
		85	-3.1	-0.9	-3.0	-0.9	-0.9	2.2

Table 7.4: Rel. bias (%) of estimated mean- imputation using stochastic BLUPs compared to deterministic BLUPs

NR model		ρ(%)	Deterministic				Stochastic			Stochastic log	
			Respmean	SL	SL+	ML	SL	SL+	ML	SL	ML
hh	MCAR	20	0.0	0.1	0.0	0.1	0.1	0.1	0.1	2.4	3.6
pers	MCAR	50	0.1	0.0	0.0	0.0	0.1	0.1	0.0	2.3	3.7
		85	0.1	0.0	0.0	0.0	0.1	0.0	0.0	2.4	4.1
hh	MCAR	20	0.7	0.0	0.2	0.0	0.0	0.2	0.0	2.3	4.0
pers	MAR	50	0.9	0.2	-0.2	-0.1	0.1	-0.2	0.0	2.5	3.7
		85	0.5	0.0	-0.2	-0.3	0.0	-0.2	-0.3	2.0	3.2
hh	MCAR	20	-2.0	-1.9	-2.1	-1.7	-1.9	-2.1	-1.7	0.3	1.4
pers	MNAR	50	-2.0	-1.9	-1.7	-1.4	-1.9	-1.7	-1.4	0.3	1.9
		85	-2.0	-1.9	-0.7	-0.7	-1.9	-0.7	-0.7	0.3	2.9
hh	MNAR	20	-1.1	-1.0	-1.3	-0.8	-1.0	-1.3	-0.8	1.2	2.4
pers	MCAR	50	-1.5	-1.4	-1.2	-0.8	-1.3	-1.1	-0.8	0.8	2.5
		85	-1.8	-1.8	-0.5	-0.5	-1.7	-0.5	-0.4	0.5	3.2
hh	MNAR	20	-3.0	-2.9	-3.3	-2.5	-2.8	-3.2	-2.5	-0.7	0.3
pers	MNAR	50	-3.1	-2.9	-2.6	-2.0	-2.9	-2.6	-2.0	-0.7	1.0
		85	-3.4	-3.2	-0.9	-0.9	-3.1	-0.9	-0.9	-1.0	2.5

maximum simulation standard error = 0.06

Appendix C - Additional tables for Chapter 6

Table 7.5: Rel. Bias (%) of estimated mean - imputation using NN hh compared to donor methods

NR model		ρ(%)	donor	class donor	hh resp	NN hh	NN hh resid
hh MCAR pers MCAR		20	-0.4	0.1	0.0	0.0	0.1
		50	0.3	0.1	0.0	0.0	0.0
		85	0.5	0.6	0.0	0.0	-0.1
hh MCAR pers MAR		20	0.3	0.3	-1.1	0.8	-0.3
		50	1.2	0.2	0.3	0.7	0.0
		85	0.9	0.4	-0.0	-0.1	-0.3
hh MCAR pers MNAR		20	-2.3	-1.8	-1.2	-2.2	-2.1
		50	-1.8	-1.8	-0.8	-1.9	-1.9
		85	-1.7	-1.4	-0.2	-1.1	-1.2
hh MNAR pers MCAR		20	-1.5	-1.0	0.1	-1.4	-1.3
		50	-1.3	-1.3	0.0	-1.3	-1.4
		85	-1.5	-1.2	0.0	-0.9	-1.0
hh MNAR pers MNAR		20	-3.3	-2.8	-1.2	-3.6	-3.4
		50	-3.0	-2.9	-0.6	-3.1	-3.1
		85	-3.0	-2.7	-0.1	-2.0	-2.1

Table 7.6: Relative Bias of population proportion (%) - donor imputation methods for Y_1 - voting preference labour and Y_2 - employed

hh	pers	random donor	class donor	NN pers	NN hh
vote labour					
MCAR	MCAR	0·0	0·3	−2·2	−1·0
MCAR	MAR	−2·0	0·2	−2·9	−0·9
MCAR	MNAR	−3·5	−3·4	−6·2	−3·9
MNAR	MCAR	−2·5	−2·7	−5·4	−2·9
MNAR	MNAR	−6·5	−6·8	−8·7	−6·2
employed					
MCAR	MCAR	−0·1	0·0	−1·3	−0·2
MCAR	MAR	−1·7	0·0	−1·6	−0·2
MCAR	MNAR	−1·9	−1·3	−2·6	−1·5
MNAR	MCAR	−1·3	−0·9	−2·2	−1·1
MNAR	MNAR	−3·2	−2·2	−3·6	−2·3

Table 7.7: Estimation accuracy for imputing *Hourly Wage Rate* - Rel. Bias (%) of estimated mean using linear models compared to donor methods

NR model		ρ(%)	class donor	NN hh	NN hh resid	SL BLUP	ML BLUP
hh MCAR pers MCAR		20	0.1	0.0	0.1	0.1	0.1
		50	0.1	0.0	0.0	0.1	0.0
		85	0.6	0.0	-0.1	0.1	0.0
hh MCAR pers MAR		20	0.3	0.8	-0.3	0.0	0.0
		50	0.2	0.7	0.0	0.1	0.0
		85	0.4	-0.1	-0.3	0.0	-0.3
hh MCAR pers MNAR		20	-1.8	-2.2	-2.1	-1.9	-1.7
		50	-1.8	-1.9	-1.9	-1.9	-1.4
		85	-1.4	-1.1	-1.2	-1.9	-0.7
hh MNAR pers MCAR		20	-1.0	-1.4	-1.3	-1.0	-0.8
		50	-1.3	-1.3	-1.4	-1.3	-0.8
		85	-1.2	-0.9	-1.0	-1.7	-0.4
hh MNAR pers MNAR		20	-2.8	-3.6	-3.4	-2.8	-2.5
		50	-2.9	-3.1	-3.1	-2.9	-2.0
		85	-2.7	-2.0	-2.1	-3.1	-0.9

Bibliography

Allison, Paul D. (1984). *Event history analysis : regression for longitudinal event data*. Beverly Hills, Calif. : Sage Publications.

Andridge, Rebecca R. and Little, Roderick J. A. (2010). "A Review of Hot Deck imputation for survey non-response". In: *International Statistical Review* 78(1), pp. 40–64.

Atrostic, B. K. et al. (2001). "Nonresponse in U.S. Government household surveys: consistent measures, recent trends, and new insights." In: *Journal of Official Statistics* 17, pp. 209–226.

Australian Bureau of Statistics (2013). *Annual Report 2012-13*. Tech. rep. Australian Bureau of Statistics.

Bankier, Michael (1999). *Experience with the New Imputation Methodology used in the 1996 Canadian Census with extensions for future censuses*. Working Paper 24. Statistics Canada.

Barroso, Lucia P, Bussab, Wilton O, and Knott, Martin (1998). "Best linear unbiased prediction in the mixed model with missing data". In: *Communs Statist. Theor. Meth.* 27.1, pp. 121–129.

Bjornstad, Jan F. (2007). "Non-Bayesian Multiple Imputation". In: *Journal of Official Statistics* 23.4, pp. 433–452.

Bodner, Todd E. (2008). "What Improves with Increased Missing Data Imputations?" In: *Structural Equation Modeling: A Multidisciplinary Journal* 15.4, pp. 651–675. DOI: 10.1080/10705510802339072.

Borgoni, Riccardo and Berrington, Ann (2013). "Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures". In: *Quality & Quantity* 47, pp. 1991–2008. DOI: 10.1007/s11135-011-9638-3.

Breslow, N.E. and Clayton, D.G. (1993). "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association*.

Brick, J. Michael (2013). "Unit nonresponse and weighting adjustment: A critical review". In: *Journal of Official Statistics* 29.3, pp. 329–353.

Bryk, Anthony S and Raudenbush, Stephen W (1992). *Hierarchical linear models: applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Buuren, Stef van and Groothuis-Oudshoorn, Karin (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3.

Carle, Adam C (2009). "Fitting multilevel models in complex survey data with design weights: Recommendations". In: *BMC Medical Research Methodology* 9.1, p. 49.

Carlin, BP and Louis, T A (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Champan and Hall.

Carpenter, James R., Goldstein, Harvey, and Kenward, Mike G. (2011). "REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types." In: *Journal of Statistical Software* 45:5. DOI: `http://www.jstatsoft.org/v45/i05`.

Chambers, Ray L. (2001). *Evaluation Criteria for Statistical Editing and Imputation*. National Statistics Methodological Series 28. University of Southampton.

Chandola, T et al. (2003). "Social inequalities in health by individual and household measures of social position in a cohort of healthy people". In: *Journal of Epidemiology & Community Health* 57, pp. 56–62.

Chen, Jiahua and Shao, Jun (2000). "Nearest Neighbor Imputation for Survey Data". In: *Journal of Official Statistics* 16.2, pp. 113–131.

Clark, Robert G and Steel, David G (2002). "The effect of using Household as a Sampling Unit". In: *International Statistical Review* 70, pp. 289–314.

Clark, Robert G. and Steel, David G. (2007). "Sampling within Households in Household Surveys". In: *Journal of the Royal Statistical Society Series A (Statistics in Society)* 170.1. URL: http://www.jstor.org/stable/4623134.

Clayton, David and Rasbash, John (1999). "Estimation in large crossed random-effect models by data augmentation". In: *Journal of the Royal Statistical Society Series A* 162.3, pp. 425–436.

Cochran, William G. (1977). *Sampling Techniques*. John Wiley & Sons.

David, Martin et al. (1986). "Alternative methods for CPS income imputation". In: *Journal of the American Statistical Association* 81.393, pp. 29–41. ISSN: 01621459. URL: http://www.jstor.org/stable/2287965.

Dempster, A. P., Laird, N. M., and Rubin, Donald B. (1977). "Maximum Likelihood from Incomplete Data via the EM algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)y* 39.1, pp. 1–38.

Di Zio, Marco and Guarnera, Ugo (2009). "Semiparametric predictive mean matching". English. In: *AStA Advances in Statistical Analysis* 93.2, pp. 175–186. ISSN: 1863-8171. DOI: 10.1007/s10182-008-0081-2. URL: http://dx.doi.org/10.1007/s10182-008-0081-2.

Diggle, Peter J. et al. (2002). *Analysis of longitudinal data*. Oxford; New York: Oxford University Press.

Durrant, Gabriele B. (2005). *Imputation Methods for Halndling Item-Nonresponse in the Social Sciences: A Methodological Review*. NCRM Methods Review Papers NCRM/002. Southampton Statistical Sciences Research Instiute, University of Southampton.

Durrant, Gabriele B. and Steele, Fiona (2009). "Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys". In: *Journal of the Royal Statistical Society Series A* 172.2, pp. 361–391.

Ecochard, Rene and Clayton David, G. (1998). "Multi-level modelling of conception in artificial insemination by donor". In: *Statistics in Medicine* 17, pp. 1137–1156.

Elbers, Chris, Lanjouw, Jean O, and Lanjouw, Peter (2003). "Micro-level estimation of poverty and inequality". In: *Econonometrica* 71.1, pp. 355–364.

Ezzati-Rice, T. M. and Khare, Meena (1994). "Modeling of response propensity in the third National Health and Nutritional Examination surveY". In: *Proceedings of Survey Research Methods Section of the American Statistical Association*, 955Ű959.

Feder, Moshe, Nathan, Gad, and Pfeffermann, Danny (2000). "Multilevel Modeling of Complex Survey Longitudinal Data With Time Varying Random Effects". In: *Survey Methodology* 26, pp. 53–65.

Fellegi, I. P. and Holt, D. (1976). "A Systematic Approach to Automatic Edit and Imputation". In: *Journal of the American Statistical Association* 71.353, pp. 17–35.

Goldstein, H. (1986). "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares". In: *Biometrika* 73.1, pp. 43–56. ISSN: 00063444. URL: http://www.jstor.org/stable/2336270.

Goldstein, Harvey (1989). "Restricted Unbiased Iterative Generalized Least-Squares Estimation". In: *Biometrika* 76, pp. 622–623.

— (1995). *Multilevel Statistical Models*. 2nd. Arnold.

Goldstein, Harvey, Healy, Michael J.R., and Rasbash, Jon (1994). "Multilevel Time Series Models with Applications to Repeated Measures Data". In: *Statistics in Medicine* 13, pp. 1643–1655.

Goldstein, Harvey et al. (2009). "Multilevel models with multivariate mixed response types". In: *Statistical Modelling* 9(3), pp. 173–197.

Gregg, Paul and Wadsworth, Jonathon (1996). "More Work in Fewer Households". In: *New Inequalities: The Changing Distribution of Income and Wealth in the United Kingdom*. Ed. by Hills, John. Cambridge University Press, p. 194.

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

Groves, Robert M. and Couper, Mick P. (1995). "Theoretical Motivation for Post-Survey Nonrseponse Adjustment in Household Surveys". In: *Journal of Official Statistics* 11.1, pp. 93–106.

Hayes, C. and Watson, N. (2009). "HILDA imputation methods". In: *HILDA Project Technical Paper Series* 2/09.

Haynes, M. A et al. (2011). "Social Determinants and Regional Disparity of Unemployment Duration in Australia: A Multilevel Approach". In: *Survey Research Conference. The University of Melbourne.*

Haziza, David (2009). "Imputation and inference in the presence of missing data". In: *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*. Ed. by Pfeffermann, Danny and Rao, C. R. Elsevier Science Publishers B. V., pp. 215–246.

Haziza, David and Beaumont, Jean Francois (2007). "On the Construction of Imputation Classes in Surveys". In: *International Statistical Review* 75.1, pp. 25–43. DOI: 10.1111/j.1751-5823.2006.00002.x.

Haziza, David and Rao, JNK (2010). "Variance estimation in two-stage cluster sampling under imputation for missing data". In: *Journal of Statistical Theory and Practice* 4.4, pp. 827–844.

Healy, J and Richardson, S (2006). *An Updated Profile of the Minimum Wage Workforce in Australia*. Tech. rep. National Institute of Labour Studies.

Healy, M.J.R. (2000). *Matrices for Statistics*. Clarendon Press.

Henderson, C. R. (1975). "Best linear unbiased estimation and prediction under a selection model". In: *Biometrics* 31.2, pp. 423–447.

Hippel, Paul T. von (2013). "Should a Normal Imputation Model Be Modified to Impute Skewed Variables?" In: *Sociological Methods and Research* 42.1, pp. 105–138.

Horton, Nicholas J. and Kleinman Ken, P. (2007). "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models". In: *The American Statistician* 61.1.

Horton, Nicholas J., Lipsitz, Stuart R., and Parzen, Michael (2003). "A Potential for Bias When Rounding in Multiple Imputation". In: *The American Statistician* 57.4, pp. 229–232. DOI: 10.1198/0003130032314.

Kalton, Graham and Kasprzyk, Daniel (1982). "Imputing for missing survey responses". In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 22–31.

Kalton, Graham and Kish, Leslie (1984). "Some efficient random imputation methods". In: *Communications in Statistics Part A - Theory and Methods* 13, 1919Ű1939. DOI: 10.1080/03610928408828805.

Laaksonen, Seppo (2002). "Traditional and New Techniques for Imputation". In: *The Journal of Statistics in Transition* 5.6, pp. 1013–1035.

— (2005). "Integrated modelling approach to imputation and discussion on imputation variance". In: United Nations Statistical Commission and Economic Comission for Europe Conference of European Statisticians May 2005. Statistics Finland.

Laird, Nan M. and Ware, James H. (1982). "Random-effects models for longitudinal data". In: *Biometrics* 38.

Landerman, Lawrence R, Land, Kenneth C, and Pieper, Carl F (1997). "An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values". In: *Sociolological Methods & Research* 26.1, pp. 3–33.

Lee, Wang-Sheng and Suardi, Sandy (2010). *Minimum Wages and Employment: Reconsidering the Use of a Time-Series Approach as an Evaluation Tool*. Tech. rep. Institute for the Study of Labor.

Leeuw, Edith D. de, Hox, Joop, and Huisman, Mark (2003). "Prevention and Treatment of Item Nonresponse". In: *Journal of Official Statistics* 19.2, pp. 153–176.

Liang, Kung-Yee and Zeger, Scott L. (1986). "Longitudinal Data Analysis Using Generalized Linear Models". In: *Biometrika* 73.1, pp. 13–22.

Lillard, Lee and Smith, James P. (1986). "What Do We Really Know about Wages? The Imporantance of Nonreporting and Census Imputation". In: *Journal of Political Economy* 94.3, pp. 489–506.

Little, R.J.A. and Su, H.L. (1989). *Item non-response in panel Surveys*. Ed. by D. Kasprzyk G. J. Duncan, G. Kalton and Singh, M. P. Wiley, New York.

Little, Roderick J. A. (1982). "Models for Nonresponse in Sample Surveys". In: *Journal of the American Statistical Association* 77.378, pp. 237–250.

Little, Roderick J A (1986a). "Missing Data in Census Bureau Surveys". In: *Bureau of the Census Second Annual Research Conference Proceedings*, pp. 442–454.

Little, Roderick J A. (1986b). "Survey nonresponse adjustments for estimates of means". In: *International Statistical Review* 54, 139Ű157.

Little, Roderick J A and Rubin, Donald B (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons. ISBN: 0-471-18386-5.

Little, Roderick J A and Schluchter, Mark D. (1985). "Maximum likelihood estimation for mixed continuous and categorical data with missing values". In: *Biometrika*, pp. 497–512.

Longford, Nicholas T. (1987). "A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects". In: *Biometrika* 74, pp. 817–27.

Marker, David A., Judkins, David R., and Wingless, Marianne (2002). "Large-Scale Imputation for Complex Surveys". In: *Survey Nonresponse*. Ed. by Groves, Robert M. et al. John Wiley & Sons Inc., New York., pp. 329–341.

Nordholt, Eric Schulte (1998). "IImputation Methods, Simulation Experiments and Practical Examples". In: *International Statistical Review* 66.2, pp. 157–180.

Ombudsman, Fair Work (2013). "Fair Work Ombudsman - Minimum Wages". In:

Patterson, H. D. and Thompson, R. (1971). "Recovery of Inter-Block Information when Block Sizes are Unequal". In: *Biometrika* 58.3, pp. 545–554. URL: `http://www.jstor.org/stable/2334389..`

Pfeffermann, Danny (1988). "The effect of sampling design and response mechanism on multivariate regression-based predictors". In: *J. Am. Statist. Ass.* 83.403, pp. 824–833.

Pfeffermann, Danny and Nathan, Gad (2001). "Imputation for Wave Nonresponse - Existing Methods and a Time Series Approach". In: *Survey Nonresponse*. Ed. by Groves, Robert M. et al. New York, USA, Wiley, pp. 417–429.

Rao, J.N.K. (1996). "On variance estimation with imputed survey data". In: *Journal of the American Statistical Association* 91, pp. 499–506.

Raudenbush, S.W. and Bryk, A.S. (1992). *Hierarchical Linear Models: Applications*. Newbury Park, CA: Sage Publishers.

Robbins, Michael W., Ghosh, Sujit K., and Habiger, Joshua D. (2013). "Imputation in High Dimensional Economic Data as Applied to the Agricultural Resource Management Survey". In: *Journal of the American Statistical Association* 108, pp. 81–95. DOI: 10.1080/01621459.2012.734158.

Rubin, D. B. (1976). "Inference and missing data". In: *Biometrika* 3.63, pp. 581–92.

— (1987a). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons.

— (1987b). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons.

Rubin, Donald B (1988). "An Overview of Multiple Imputation". In: *Proceedings of the survey research methods section of the American statistical association*, pp. 79–84.

Rubin, Donald B. (1996). "Multiple Imputation After 18+ Years". In: *Journal of the American Statistical Association* 91, pp. 473–489.

Sande, I. G. (1983). "Incomplete Data in Sample Surveys". In: ed. by Madow, William Gregory, Nisselson, Harold, and Olkin, Ingram. Vol. 3. New York: Academic Press. Chap. Hot-deck imputation procedures, pp. 339–349.

Sarndal and Lundstrom (2005). "Estimation in Surveys with Nonresponse." In: Jonn Wiley & Sons Ltd. Chap. Selecting the Most Relevant Auxiliary Information, p. 110.

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

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*. John Wiley & Sons., p. 506.

Shao, Jun (2000). "Cold deck and ratio imputation". In: *Survey Methodology* 26.1, pp. 79–85.

— (2007). "Handling survey nonresponse in cluster sampling". In: *Survey Methodology* 33.1, p. 81.

Singh A.C. Grau, E.A. and Folsom, Jr. R.E. (2001). "Predictive mean neighborhbourhood imputation with application to the person-pair data of the national household survey on drug abuse." In: *Proceedings of the Annual Meeting of the American Statistical Association,*

Skinner, Chris J and D'Arrigo, J (2011). "Inverse probability weighting for clustered nonresponse". In: *Biometrika* 98.4, pp. 953–966.

Tanner, Martin A. and Wong, Wing Hung (1987). "The calculation of posterior distributions by data augmentation". In: *Journal of the American Statistical Association* 82.398, pp. 528–550.

Taylor, Marcia Freed et al., eds. (2010). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices.* http://www.iser.essex.ac.uk/ulsc/bhps/doc/.

Wang, Jichuan, Xie, Haiyi, and Fisher, James H. (2012). *Multilevel Models: Applications Using SAS.* Higher Education Press and Walter De Gruyter, p. 24.

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.

Watson, Nicole (2007). "Using Imputed Data: Examples from the HILDA Survey". In: *The Australian Economic Review* 40.4, pp. 453–61.

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

White, Ian R., Royston, Patrick, and Wood, Angela M. (2009). "Multiple imputation using chained equations: Issues and guidance for practice". In: *Statistics in Medicine* 30.4, pp. 377–399. DOI: DOI:10.1002/sim.4067.

Williams., Todd.R. and Bailey, Leroy (1996). "Compensating for missing wave data in the Survey of Income and Program Participation (SIPP)". In: *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 305–310.

Wun, Lap-Ming et al. (2007). "On modelling response propensity for dwelling unit (DU) level non-response adjustment in the Medical Expenditure Panel Survey (MEPS)". In: *Statistics in Medicine* 26, pp. 1875–1884.

Yan, Ting, Curtin, Richard, and Jans, Matthew (2010). "Trends in Income Nonresponse over Two Decades". In: *Journal of Official Statistics* 26.1, pp. 145–164.

Yuan, Ying and Little, Roderick JA (2007). "Parametric and Semiparametric Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Item Nonresponse". In: *Biometrics* 63.4, pp. 1172–1180.

Yucel, Recai M. (2008). "Multiple imputaiton inference for multivariate multilevel continuous data with ignorable non-response". In: *Philosophical Transactions of The Royal Society A* 366, pp. 2389–2403.

— (2011). "Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data". In: *Statistical Modelling* 11, pp. 351–370. DOI: 10.1177/1471082X1001100404.

Zeger, Scott L. and Liang, Kung-Yee (1986). "Longitudinal Data Analysis for Discrete and Continuous Outcomes". In: *Biometrics* 42, pp. 121–130.