Understanding the effects of weighting on parameter estimation for multilevel models

Olusegun Afis Ismail, BSc, MSc.

Submitted for the degree of Doctor of Philosophy
Applied Social Statistics
Lancaster University

September 2020

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Olusegun Afis Ismail

Abstract

Creating a model from hierarchical data with missing data without addressing the missingness in the data may lead to poor parameter estimates. Hence the overall aim of this thesis is to investigate the effects of weighting adjustment methods on parameter estimates of weighted multilevel linear model to address unequal sampling selection and non -response in the continuous response variable. The significance of this thesis is that it seeks to fill the gaps in the existing body of work on complex survey data analysis on the identification of best weighting adjustment and conditions suitable to achieve reliable estimates from a weighted multilevel linear model.

To achieve the aim of this thesis, the Jamaica Survey of Living Conditions 2007 was used in a simulation study on a weighted multilevel linear model of the annual household expenditure of meals purchased away from home to investigate four weighting adjustments methods using two scenarios: Missing at Random (MAR) and Missing not at Random (MNAR) at 20%, 40% & 60% rate of missing respectively in the outcome variable. In order to fully investigate the effects of the weighting adjustments, the missingness in the outcome variable were tested as a function of a continuous, categorical and combination of both continuous and categorical variables. The simulation study was also extended to the scaling of the weights to identify changes in the effects on the parameter estimates.

The weighting adjustment with the most reliable estimates were applied to the modelling of reported income from the China Health and Nutrition Survey (CHNS 1989 & 2011) data as well as the household expenditure of meals purchased away from home in the Jamaica Survey of Living Conditions 2007 respectively. In the application of the weighting adjustments, different scenarios on different multilevel models were investigated to identify any changes in the parameter estimates.

The findings from the simulation study on the random effect multilevel model revealed that sampling weight adjusted for missing data using item non-response weight produced the most reliable estimates of the fixed component of the linear multilevel models at the 20% rate of missing.

Acknowledgements

I would like to express my sincere gratitude to Professor Peter Neal who came on board, supervised, and provided guidance that made this journey memorable. I would also like to thank Dr. Novie Younger-Coleman for her supervision, ideas and words of encouragement over the years. I would like to extend special gratitude to the late Professor Damon Berridge and Dr. Gillian Lancaster for their supervision at the beginning of my journey at Lancaster to the point of their departure. Special thanks to Dr. Kanchan Mukherjee for addressing all the challenges faced during my studies at Lancaster. Special thanks to my employer, the University of Technology, Jamaica for offering me a scholarship to embark on this Journey which has sustained me over the years. I would also like to thank my colleagues at the School of Mathematics and Statistics, Faculty of Science and Sports, University of Technology, Jamaica especially, Mr. Kevin Williams, Mrs. Ave McIntosh, Mr. David Finlay, Mr. Gustwell Weir, Mr. Leslie Young, and Mr. Errol Rowe for their understanding and support. I would also like to express thank you to Mr. Frederick Gordon of the Planning Institute of Jamaica (PIOJ), Mr. Douglas Forbes, former Director of Surveys at the Statistical Institute of Jamaica (STATINJA) and Ms Revene Evans for redrawing the maps.

Finally, special thanks to my family and friends for their love, support, encouragement and enduring with me when I was missing in action busy studying. And to Almighty God for giving me the strength to complete this journey.

Dedicated to

My Family

Contents

Chapter 2 Literature Review 6

Chapter Outline

Chapter 3 Exploratory Data Analysis 17

Chapter Outline

List of Figures

List of Tables

Abbreviations

CHNS	China Health and Nutrition Survey
CHNS1989	China Health and Nutrition Survey 1989
CHNS2011	China Health and Nutrition Survey 2011
ED	Enumeration District
JSLC	Jamaica Survey of Living Conditions
JSLC2007	Jamaica Survey of Living Conditions2007
HS	Household selection weight
IR	Item non-response weight
MAR	Missing at Random
MNAR	Missing not at Random
PS	ED selection weight
PHS	ED and Household selection weight
STATIN	Statistical Institute of Jamaica
SWAM	Sampling Weight Adjusted for Item non response

Chapter 1

Introduction

1.1 Overview of Survey Data Sampling History

The methods of sampling in surveys have been found to be an important in information gathering about a population (Skinner & Wakefield, 2017). The theory behind sampling can be traced to the papers presented by the early statisticians such as Nayman (1934); Hansen and Hurwitz (1934); Sukhatme (1935); and Bowley (1936). These studies enable further development on sampling theory in the nineteen forties and early fifties (Cochran, 1942; Madow & Madow, 1944; Horvitz & Thompson, 1952). These developments on sampling methods continued to enhance the estimates from sample survey about the population. However, the data collected from surveys used to derive these estimates are not immune from non-response problems which are ubiquitous in every survey data collected either via simple or complex sampling techniques (Hansen & Hurwitz, 1946).

1.2 Census and Surveys in Jamaica

The Statistical Institute of Jamaica (STATIN), a Government Agency, has the responsibility for conducting census and household surveys in the island in order to collect national information and country statistics. Since its inception in 1946, the Agency has produced publications such as the consumer price indices reports, demographic statistics, employment and income levels, reports on the environmental statistics and state of the environment report, external trade reports, labour force statistics, national income and product, pocketbook of statistics, population census reports, production statistics, quarterly gross domestic product and statistical reviews (www.statinja.gov.jm).

In order to monitor the demographic dynamics and statistics of the population, STATIN conducts a census every ten years. The thirteenth census was conducted in 2001 and the fourteenth was conducted in 2011. Historically, the first census was conducted in 1844, and the Act of Parliament for Population and Housing Census was passed in July 30, 2001 (Population Census, 2001). The Population and Housing census provide information on the demographics and housing spread of the population. It also serves as the population reference for researchers in selecting samples from the population.

1.3 Role of Complex Survey in Policy Development

In developed and developing countries, complex surveys are non-uniform in the sampling design and are becoming a popular tool for data collection in policy formulation. In Jamaica, the Statistical Institute of Jamaica (STATINJA) has the mandate to collect data to inform policy development and the Institute is noted for a series of complex surveys. The Labour Force Survey (LFS) is one of such surveys. The Labour Force Survey publications usually informed the Government of Jamaica on employment and unemployment rate on quarterly and annual basis. This publication enables the Ministry of Labour to formulate policy that guide labour matters such as the issuing of work permit and job opportunities. The Jamaica Survey of Living Conditions (JSLC) is another complex survey that is published by the Planning Institute of Jamaica while the Statistical Institute of Jamaica is responsible for the data collection. The JSLC is published on an annual basis to inform the Government of Jamaica on the prevalence of poverty. The JSLC enables the Ministry of Labour and Social Security to adequately plan for the beneficiaries on the Programme of Advancement through Health and Education (PATH). The STATINJA also collect data for the Household Expenditure Survey (HES) used to produce the consumer price index on an annual basis. In Jamaica, other institutions seldom conduct complex surveys for decision making, the Jamaica Life Style Survey is an example of such survey conducted by the Caribbean Health Institute within the University of the West Indies. The University of Technology, Jamaica also conduct Adult Population Survey for the Global Entrepreneurship Monitor to measure nascent entrepreneurs and other type of entrepreneurial activities among Jamaica adult population.

In North America especially the United States and Canada, there are a range of complex surveys that inform the policy of the Government in both Countries. In the United States, the centre for disease control is known for the National Health and Nutrition Examination Survey (NHANES) which enables the American Government to monitor the nutrition and health status of American citizens inclusive of sub populations with the USA.

The U.S. Census Bureau, the Government Agency responsible for census, household and business surveys and often use a multi-stage approach in the sample design. The Statistics Canada also conducts series of surveys that informs the policy of the Canadian Government. The Statistics Canada Labour Force Survey is another example of a complex survey.

1.4 Aim and objectives of the thesis

The aim of this thesis is to assess weighting adjustments from different scenarios in multilevel models of continuous outcomes with the ultimate view of identifying the least bias weighting adjustment for compensating for unequal sample selection and missingness due to non-response problems. The continuous variable in the complex survey data of interest is the annual household expenditure of meals purchased away from home in the JSLC2007 and the reported income in the CHNS 1989 and 2011. This thesis seeks to achieve four objectives relating to weighting methodological issues especially when compensating for unequal sample selection and missing data due to non-response in the modelling of a continuous outcome. The first objective is to develop a series of weighting methods that can address unequal sample selection in design of complex surveys and missing data due to non-response problems. The second objective is to conduct exploratory data analysis of the JSLC2007 and CHNS 1989 and 2011 in preparation for building a series of unweighted and weighted linear multilevel models using continuous outcome variable to assess the prediction of the variable in the presence of missing data and unequal sample selection in the complex survey data. These models will also incorporate survey design information to enable comparisons of parameter estimates. The third objective is to conduct a simulation study on weighted multilevel models so as to identify the best weighting adjustment method by comparing parameter estimates to assess the bias from the weighting methods. The fourth objective seek to illustrate the application of the best weighting adjustment method to the analysis of the JSLC2007 and CHNS 1989 and 2011.

The goal of this research is to investigate the strength and limitations of various weighting approaches for addressing unequal sample selection and missing data issues that arise because of non-response when modeling a continuous outcome variable in complex survey data.

1.5 Study research questions

The study objectives were developed to address the following questions:

1. Which weighting adjustments is the least bias when addressing unequal sample selection in complex survey design and missing data problem due to item nonresponse from survey respondents?

2. What are the conditions such as the proportion of missing and missing mechanism which the least bias estimates occur for the identified weighting adjustment method?

3. What are the effects of Intraclass Correlation Coefficient (ICC) on different parameter estimation methods?

4. Does the number of clusters in a level 1 or level 2 units of unweighted or weighted linear multilevel model affects the final predictors in the model?

1.6 Thesis structure

This thesis contains seven chapters. The Literature Review on complex survey and associated issues are provided in Chapter 2. Exploratory analysis of the Jamaica Survey of Living Conditions (JSLC2007) and China Health and Nutrition Survey data are provided in Chapter 3. Chapter 4 focuses on the statistical methodology used in the thesis. The methodology includes detailed explanation of the different weighting adjustments, parameter estimator for the linear mixed effects models and logistics regression model. The simulation study for the multilevel linear model and the incorporation of survey information are provided in Chapter 5. Chapter 6 application of weighting adjustment method to Jamaica Survey of Living Conditions and China Health and Nutrition Survey. Chapter 7 contains discussions, conclusions and future work.

Chapter 2

Literature Review

Chapter Outline

In Section 2.1, a review of what constitutes a complex survey and complex survey data is presented. In Section 2.2, the reviews of non-response, causes and associated problems are presented. In Sections 2.3 and 2.4, the reviews of unequal probability of sample selection in complex sampling design and how to address missing data are presented respectively. In Sections 2.5, 2.6, 2.7 and 2.8 the reviews of weighting adjustment approaches viz-a-vis design-based or model-based, type of weighting adjustments, multilevel modeling in relation to cluster size and intraclass correlation coefficient, and imputation are presented respectively. -And finally, Section 2.9 contains the major contribution of this thesis to the existing body of knowledge.

2.1 Understanding Complex Survey and Complex Survey Data

There is a momentum on the analysis of complex survey data in the literature. However, this momentum has led to the need for more clarifications on several issues surrounding complex sample design and the associated problems. Nearly every discipline now conducts studies involving data with the characteristics of a complex survey and in some instances, an analytical inference in the form of multilevel modelling.

In survey analysis related studies, the word "complex survey" often refers to a survey in which the sample designs involve some form of clustering and/or stratification in stages of the sample selection (Carle, 2009; Skinner and Wakefield, 2017). A complex survey may be characterized by: unequal probability in the sample selection, multi-stages in the sample selection, clustering, stratification, non-response issues and confidentiality of the survey respondents (Lee and Forthofer, 2006; Skinner, Holt and Smith,1989; Hansen and Hurwitz, 1943; Pfeffermann, 2011; Heeringa et.al, 2017). These features have made surveys with complex sampling useful in obtaining a representative sample of "hidden and hard -to-reach population" as cited in Wirth and Tchetgen (2014).

The highlighted characteristics of a complex survey are also found in large surveys conducted by National Statistical Offices and some Private Institutions that requires a nation-wide survey. These features differentiate a complex survey from other type of surveys, especially when they are household-based, nation-wide and requires some form of nesting/hierarchy in the sampling structure. A typical example, is the Jamaica Survey of Living Conditions (JSLC) series, which is key data set of interest in this thesis. Similar surveys with complex sample designs include the National

Health and Nutrition Examination Survey (NHANES, 2000) and National Health Interview Survey (NHIS) in the United States of America (USA). Other notable surveys with these features include the Family Expenditure Survey (FES) in the United Kingdom and the China Health and Nutrition Survey conducted in China by the University of North Carolina, USA. According to Johnson and Elliot (1998), conducting these types of surveys often support efficient use of resources and may be one of the reasons for the preference by the National Statistical Offices and some Private Institutions. The United Nations Guidelines on Household Surveys also supported the use of complex survey design (UN, 2005) for nationwide household-based surveys. In reference to the highlighted examples, the sampling design is usually in stages. Furthermore, the clustering of the target population may be followed by the stratification or vice versa as the case may be in some preferred approach during the sampling processes. For example, Kish (1965, pg. 359) demonstrates how complex survey designs can be constructed in practice using the United States. Dwellings reside within counties which can be used to construct complex surveys at a state or national level.

2.2 Nonresponse, Causes and Associated Problems

In survey sampling, Kish (1995, pg.532) defines non-response as many sources of failure to obtain response or measurement on selected sample. This definition is comparable to all the non-response definitions found in the literature (Bethlehem, 2009, pg. 1, Singer 2006). The non-response is viewed as a major problem facing researchers in social and medical sciences and Official Statistics (Durrant and Steele, 2013). This problem is worldwide that Durrant and Steele (2013) further cited the decreasing response rate for surveys from Martin and Matheson (1999) and also elsewhere according to Dette (1999) and Steeh et.al (2001). Singer (2006) cited Groves and Couper (1998) to affirmed that the aspect of nonrespondent in a survey that are of greatest concern to survey methodologists are noncontact and refusal. Singer (2006) further trace the history on early research on non -response to polling in the 1930s and stated that first non-response article in JSTOR statistical journals was from 1945 and such article in the Public Opinion Quarterly was from 1948. Singer (2006) concluded by stating that “one of the important scientific challenges facing survey methodology at the beginning of this century is determining the circumstances under which nonresponse damages inferences to the target population”. Singer (2006) also highlighted the second challenge which “is the identification of methods to alter the estimation process in the face of nonresponse to improve the quality of the sample statistics”. In the attempt to study non-response rates and non-response bias in household survey, Groves (2006) expressed non-response bias as the departure of the expected value from an estimate from its true value.

The two categories of non-response in the literature are unit and item non-response (Lahor,2009, pg.329; Rao,2000 pg.216). Kish (1995, pg.532) outline causes which may lead to unit non-response such as not-at-home situation, out right refusal, incapacity or inability due to illness, not found, and lost schedules. In addition to the list from Kish (1995), non-response can also be due to poorly worded survey items or poor interviewing skills in conducting a survey (Scheaffer et.al, 1996). Thus, creating partial or total missing data and departure from the expected response rate that may ultimately lead to a bias in the sample estimate if not address. In this study, the Unit non-response is not contemplated for the fact that the profile information of the respondents in the sample frame and reasons for not participating in the survey of the respondents are not available to secondary users. These highlighted issues clearly support continued investigation on non-response especially on the context of rate of missing from non-response causes with regards to hierarchical modeling which will complement the highlighted reviews.

2.3 Unequal Probability of Sample Selection in Complex Sampling Design

In complex survey, the sampling processes are not uniform as stated earlier on the characteristics. Owing to the multistage in the sampling schemes, selected samples often have unequal probabilities of selection at some or all stages of the sampling process (Pfeffermann et.al,1998). These unequal selection probabilities at any stage of the sampling in complex survey may induce bias in standard estimation in the development of multilevel models (Pfeffermann et.al,1998). Despite the views from Pfeffermann et.al, (1998) not all studies relating to hierarchical modelling embrace the adjustment of estimates for unequal selection probabilities as demonstrated in Farrell and Ludwig (2008). For some analysts from other disciplines, the accuracy of estimates in the use of hierarchical maximum likelihood estimates compared with parameter estimation from classical single -level model serves as the motivation besides the adjustment for unequal sample selection in the data. This is highlighted the field of psychology by Rouder et.al (2005). The need for continuing investigation on the importance of incorporation of weight adjustments to address the unequal sample selection probabilities to improve parameters estimation from multilevel model cannot be overemphasize especially for non -survey methodologist.

However, survey statisticians suggested that weighting adjustments should be used to address the bias in the estimates that is due to unequal selection probability in statistical inference (DuMouchel and Duncan, 1983; Pfeffermann,1993). This suggestion did not provide comparisons for each adjustment proposed in the sampling process to ascertain at which adjustment and what conditions in terms of missingness do these adjustments reduced or eliminate the bias in the estimate.

2.4 Addressing Missing Data Problems

To address the missing data problem in the post data collection phase, two methods are widely cited in the literature, this includes weighting adjustments and imputation (Little & Rubin, 2002). Furthermore, use of weighting adjustment methods for non-response problems in the context of weighted model estimates using complex survey data are well documented in Holt and Elliot (1991), Himelein (2013); Little and Rubin (2002); Korn and Graubard (1995); Pfefermann (1993); Kish and Frankel (1974); Hansen and Hurwitz (1946). However, findings from these studies usually provide comparisons between weighted and unweighted estimates in most instances along with the type of estimators but were found to be limited in terms of comparative analysis on the percentage of missing data for the weighted methods in which these methods are effective.

2.5 Context of Weighting Adjustment Approaches: Design or Model -Based

The context in which weighting adjustment methods are used in parameter estimation from complex survey data are yet to be fully explored and has led to a continuing debate on the types of estimation methodology, especially in hierarchical modelling in the context of design-based or model-based approaches (Snijders and Bosker, pg.217, 2012). According to Skinner and Wakefield (2017), the model-based approach is common in the main stream statistical curriculum while the design-based approach may require specialized training which continue to gain momentum but has been mostly limited to survey designers and practitioners. The focus of this thesis is the application of weighting adjustments in the context of hierarchical model which is purely a design-based approach. This approach will enable further investigation on the suitable conditions for weighting adjustments in the inference for multilevel models.

2.6 Understanding Weighting Adjustments

There are several approaches to weighting in the literature. These approaches continue to generate further investigation depending on the context and purpose of the adjustment of the estimates. Of all the weighting adjustment approaches, the post stratification is the most popular in practice for adjusting estimates for population total using known auxiliary variables (Little, 1993; Lohr,2010; Rao, 2000). The post stratification approach is common practice by the National Statistical Offices (NSO) as a tool to post stratify estimates to reflect population figures. In the literature, the post stratification weight continues to be identified with different names and under different context with slight modifications. In some studies, post stratification is described in the context of "raking", "weighting class adjustment" and "calibration", these terms have been used interchangeably (Holt and Smith, 1979; Smith, 1991; Deville and Sarndal, 1992; Deville, Sarndal & Sautory, 1993; Kalton and Flores-Cervantes, 2003; Little & Rubin, 2002; Valliant, Dever & Kreuter, 2013). In most cases, the context of use of post stratification adjustments is to address the unit non-response (Heziza & Lesage, 2016). However, some difficulty may arise whenever the census population figures or needed information for which the proportion for homogenous group is to be created and this may create challenges to researchers who may want to create post stratification weight.

Unlike the weighting adjustment by post stratification that requires an external data in most cases, the weighting adjustment for the item non-response often uses the available data in the adjustment and this also enable the application of Little & Rubin (pg.19, 2002) missing mechanisms. The common item non-response weight in the literature is usually created by the inverse response probability developed through a logistic model of response variable (Skinner and Arrigo, 2011, Bethlehem et.al, 2011, Barbara and Stephen, 2001). Understanding of the performance of the item non-response weight under different rate of missing cases is limited in the literature. Available studies often neglect to state what percentage of missing cases the weight is being applied.

From the highlighted weighting adjustments, the focus of this study shall be on the theory of weighting adjustments to address unequal sample selection due to differential responses in the sampling units and also item non-response. Similarly, these weights will be incorporated in the estimation of the parameter to highlight the effect of the weight in the estimation process.

2.7 Multilevel Modelling: The Number of Cluster and Intraclass Correlation

There are several issues surrounding the application of multilevel modelling in real life situations. One of the issues is the use of the multilevel modelling when the intraclass correlation is very small without proper definition of what is considered small. Another problem is the number of clusters in the model. These have led to numerous studies on these issues. Based on observations and also some literatures reviewed, the majority of these concerns are emanating from users of statistics especially the field of psychology and sociology to a greater extent compared to the perspective of survey statisticians. For example, McNeish and Stepleton (2014) conducted a review on previous studies on multilevel models with small number of clusters and also provided an illustrative simulation to demonstrate how a simple model becomes adversely affected by small clusters. McNeish and Stepleton (2014) cited Kreft (1996) who postulated a cluster size of 30, also cited Snijders and Bosker (2012) who recommended 20 clusters and Hox (1998, 2010) recommending 50 to 100 clusters. McNeish and Stepleton (2014) claimed that a review of 13 journals from education, psychology and sociology did not meet these recommendations. The findings of McNeish and Stepleton (2014) provided minimum number of clusters varying from 5 to 50 for continuous outcomes with indication consideration for the adjustment of unequal probabilities of selections of the clusters in the model nor adjustment for missing data.

Some researchers such as Huang(2018) highlighted the circumstances surrounding intraclass correlation and the justification of multilevel modelling. Huang(2018) cited Nezlek (2008, p.856); Hayes (2006); and Thomas and Heck (2001) as studies spreading the myth that 'when the intraclass correlation is low, multilevel modelling is not needed" . In this study one of the objectives is to investigate the changes in the predictor variables as the intraclass correlation coefficient changes to see if there is any truth in this myth. Similarly, another objective is to investigate the effect number clusters on the predictor variables to identify any possible changes as the cluster numbers varies in the level 1 and level 2 units to provide solutions and recommendations from the perspective survey statistician.

2.8 Imputation as an Alternative to Weighting Adjustments

The use of imputation technique to address missing data in statistical analysis has been around for a long time. The use of imputation method is also common in the literature as a tool for addressing item non-response problem, especially the hot and cold deck imputation, as well as the substitution of the mean and multiple imputation (Little and Smith,1987; Rubin and Little, 2002; Andridge and Little, 2010; Lohr, 2010; Little & Rubin, 2002; Rao, 2000; Buuren, 2012; Schafer & Schenker 2000).

According to Buuren (2012) the statistical method to replace a missing value were developed by Allan and Wishart (1930) this clearly suggest how long imputation method had been in existence. Buuren (2012) further posited that the methods were further enhanced by Yates (1933) and others such as Dempster et.al, 1977; Madow et.al., 1983; Little and Rubin (1987). The work of Little and Rubin (1987) is now recognized as the leading procedure for imputation techniques. Little and Rubin (2002) described imputations as “means or draws from a predictive distribution of the missing values that requires a method of creating a predictive distribution for the imputation based on the observed data”.

There are two popular approaches to generating predictive distribution for imputation that is based on the observed data (Little and Rubin, 2002). This includes explicit and implicit modelling approaches. According to Little and Rubin (2002), the implicit methods include the hot deck imputation, substitution, cold deck imputation and a composite approach. On the contrary, the explicit method includes mean imputation, regression imputation and stochastic regression imputation.

According to Tabachnick and Fidell (2000), the mean substitution form of single imputation has one major disadvantage which is the decrease in variability between individuals’ responses and biases correlations. In some National Statistical Offices (NSO) such as the Census Bureau in the USA, the hot deck single imputation is still in practice (Reilly & Pepe, 1997) as cited in Patrician (2002), despite the disadvantages. The advances in the imputation method have led to multiple imputations over single imputation when the idea started. The multiple imputation method was developed to address the disadvantages in the single imputation method (Patrician, 2002). Buuren (2012) attributed the multiple imputation to Rubin’s work in the 1970s and continued to gain momentum over single imputation with the work of Allison (2000) and Schafer (1997, 2000).

Whereas the multiple imputation is very straight forward for addressing missing data in non- hierchical data this is not the case for the clustered data because the software involved and need for advanced statistical knowledge (Grund, Ludtke and Robitzsch, 2016). Available software programmes are based on different procedures. According to Grund et.al (2018) the procedures can be grouped into two broad paradigms: the joint modelling approach (JM) and fully conditional specification (FCS) approach. For the JM approach, a single model is specified for all variables with missing data and imputations are simultaneously generated from this model for all variables with missing data (Grund et. al, 2018). This is unlike the FCS approach where data are imputed separately for each variable with missing data while conditioning on some or all the other variables in the data set (Grund et.al, 2018). However, Grund et.al (2018) further stated despite the difference in approach, both approaches use similar covariance structures at the individual and group level and can be used interchangeably as demonstrated by Carpenter & Kenward (2013), p.220; Ludtke et.al, 2017; and Mistler (2015). These approaches can be found in software packages such as Mplus and in R packages jomo (Quartagno & Carpenter, 2016). In summary, Multiple Imputation can be explored using any of these approaches in the software packages to address missing data in multilevel data and also for comparison purposes with estimates from weight adjustment methods.

2.9 Major Contribution of the Study

The four main contributions of this study to the statistical analysis of weighting adjustment methodology. The first contribution is the identification of weighting adjustments that best address bias in estimates when there are missing data that is due to nonresponse specifically item non-response and unequal sampling selection when developing hierarchical model with a linear continuous response. The second contribution is on the investigation of the limitation of the weighting adjustments under varying percentages of missingness in the data. The third contribution is a comparative analysis of results from models from small number of clusters versus large number of clusters. The fourth and final contribution is the comparative analysis of results of hierarchical models with imputed data and weighted adjustments for item nonresponse to guide future developments in survey data analytical inference.

Chapter 3

Exploratory Data Analysis

Chapter Outline

This Chapter seeks to provide the exploratory analysis of selected variables from the Jamaica Survey of Living Conditions 2007(JSLC2007) and China Health and Nutrition Surveys for the year 2009 and 2011 (CHNS2009 and CHNS2011) respectively. The exploratory analysis will serve as the preliminary understanding of the study variables prior to the multilevel modelling activities in Chapter 5. It will also provide an insight on the degree of relationships between these variables from each population. The first section in the Chapter provides a brief history of Jamaica then proceed to the second section on Jamaica Survey of Living Conditions (JSLC) and associated sub-sections. The follow-up sections are devoted to the exploratory analysis of the secondary data: CHNS2009 and CHNS2011. The Chapter is concluded with the summary of the exploratory analysis.

3.1 Brief History of Jamaica

The island of Jamaica is divided into fourteen (14) parishes. Each of the parishes has an administrative capital, except for Kingston and St. Andrew which has joint administrative capital, downtown Kingston and a Mayor responsible for both parishes. Further to the census of 2001 for which the survey of living conditions of 2007 sample was drawn, the parish of Kingston (100%-urban; Table 3.1) had the highest proportion of urban Enumeration Districts (EDs) among the parishes followed by the parish of St. Andrew (88% -urban; Table 3.1).

Some of the manufacturing companies and banks had their headquarters in the City of Kingston or the greater area known as the Kingston Metropolitan Area (KMA) which includes the urban area of St. Andrew. As a result of the proximity of the parish of St. Catherine to both Kingston and St. Andrew, urbanization of the parish has seen rapid increase especially the City of Portmore. The population for the parish of St. Catherine grew by 7 percent while that of St. James grew by 5 percent between 2001 and 2011 census (2011 Census of Population and Housing – Table 3.1). However, there are also increases in the number of urban districts for some parishes and decrease in some instances. The contrast is evident in the distribution of the number of rural and urban districts boundaries for 2011 Census (Table 3.1). There are also two international airports which significantly contributed to the development of Kingston and the parish of St. James. There are more tertiary and non-tertiary institutions in the KMA compared to the other parishes as well as the opportunities for employment. The KMA comprises of the parishes of Kingston and St. Andrew. The parish of St. Andrew has two public universities and two private universities, which also accounts for 21.25 percent of the island's population while Kingston accounts for 3.30 percent (2011 Census of Population and Housing). These combined factors, places the population of the KMA to 24.55 percent. The lower prevalence of poverty among the household sampled in the Survey of Living Conditions for 2007 could likely be attributed to the employment and access to education in the KMA.

High levels of poverty are widespread in the parishes with high proportions of rural enumeration districts (Tables 3.1). These parishes also had less tertiary type institutions as illustrated in Figure 3.1 but are dominated by mineral extraction and agricultural type activities (Clarendon, St. Catherine, St. Thomas and Trelawny). It is noteworthy to state that the map in Figure 3.1 was drawn in ArcGIS 9.2 after collating the number of tertiary institutions within each Parish. The parishes of St. James, St. Ann and Westmoreland are renowned for tourism. In 2007, agriculture, forestry and fishing contributed 5 percent to the island gross domestic product when compared with the goods producing industry and mining and quarrying of natural resource, a leading contributor to the gross domestic product, 8.5 percent. The gains from the export of natural resources (for example bauxite) and brown sugar have suffered significantly from the continued rise in the cost of crude oil which is the main component in the energy requirement for the semi processing of the bauxite and sugar cane. Figure 3.2 illustrates the island development by enumeration districts (ED).

Figure 3.1: Spread of the registered tertiary institutions in Jamaica by the University Council of Jamaica

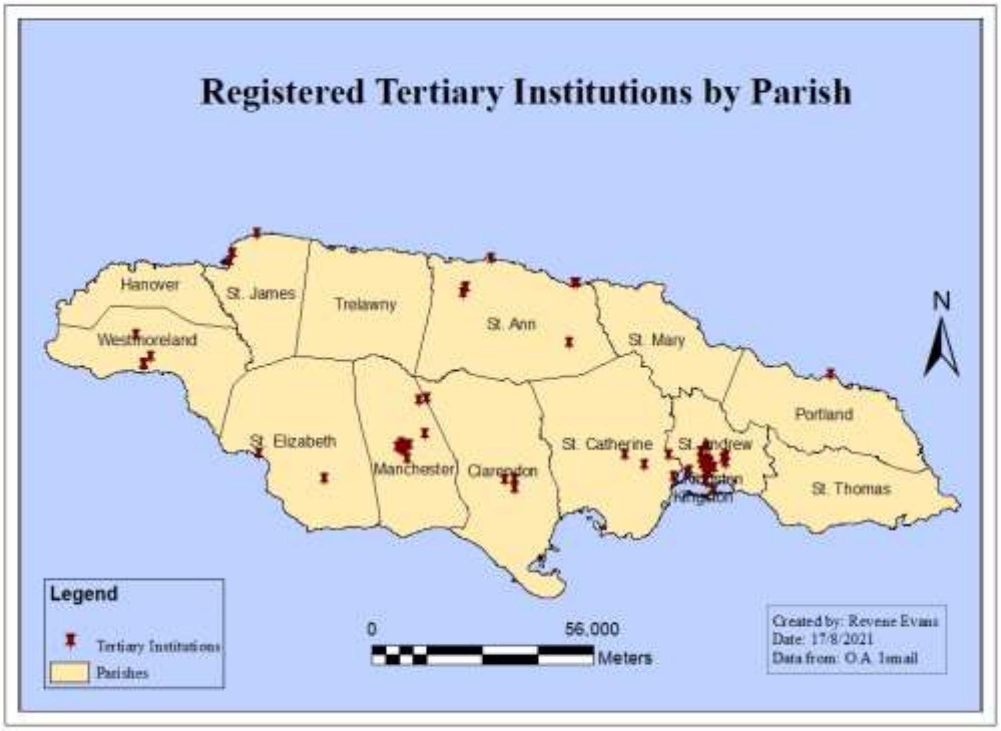
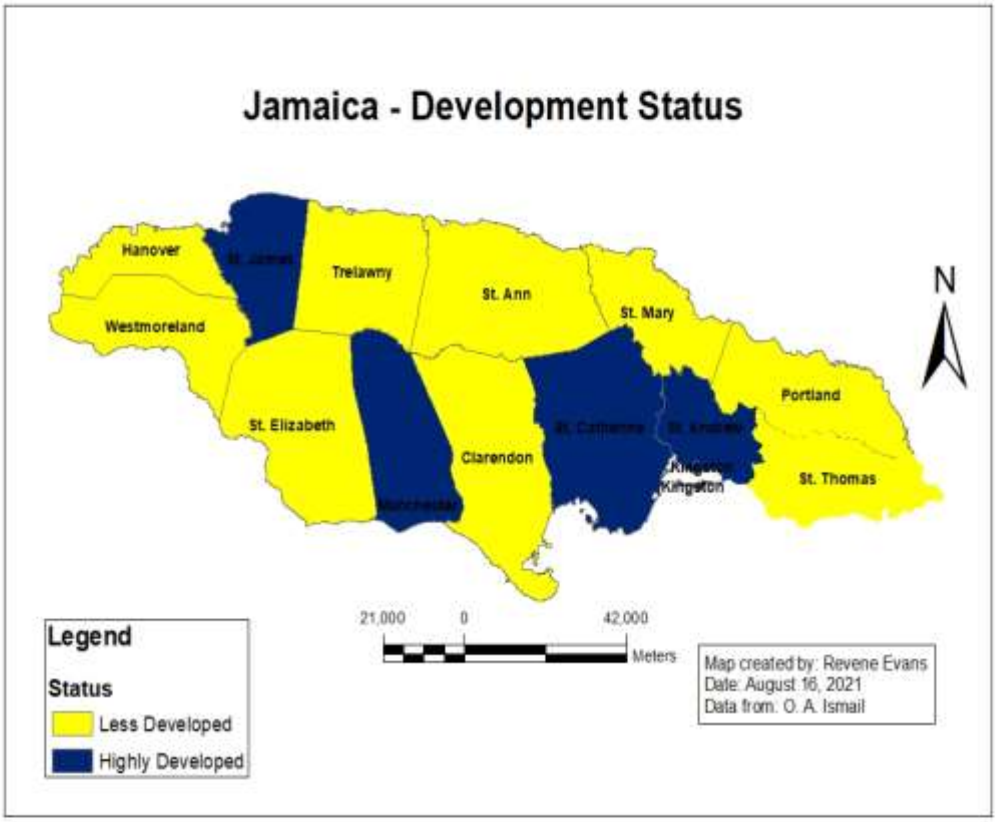


Table 3.1: Percentage (%) spread of Enumeration Districts (ED) by Parish for Census 2001 & 2011

		2001		2011	
		Urban %	Rural %	Urban %	Rural %
Parish	**Capital**	ED	ED	ED	ED
Kingston		100	0	100	0
St. Andrew	**KMA**	88	12	87	13
St. Thomas	**Morant-Bay**	25	75	28	72
Portland	**Bull- Bay**	22	78	24	76
St. Mary	**Port-Maria**	19	81	24	76
St. Ann	**Saints-Ann's Bay**	24	76	27	73
Trelawny	**Falmouth**	16	84	20	80
St. James	**Montego-Bay**	57	43	56	44
Hanover	**Negril**	10	90	11	89
Westmoreland	**Savana-lar-mar**	22	78	25	75
St. Elizabeth	**Black River**	13	87	15	85
Manchester	**Mandeville**	32	68	33	67
Clarendon	**May Pen**	28	72	31	69
St. Catherine	**Spanish Town**	70	30	73	27

Figure 3.2: Spread of the Enumeration Districts to illustrate the level of development by percentage of urban and rural areas


Jamaica - Development Status

Map created by: Revene Evans
Date: August 16, 2021
Data from: O. A. Ismail

3.2 Jamaica Survey of Living Conditions (JSLC)

The Government of Jamaica had introduced the Survey of Living Conditions in 1988 to assess the living conditions of the Jamaican population, in order to develop social programmes (JSLC, 1988). Since 1988, several surveys had been conducted to continuously understand the social problems facing the population which had led to the development of several initiatives by the Government. A notable initiative is the Programme of Advancement Through Health and Education (PATH). The Government of Jamaica has also established the Social Development Commission (SDC) and the Jamaica Social Investment Fund (JSIF).

This study uses the data from the survey which was conducted in 2007. It is worth mentioning that similar analyses could have been conducted for data from other years. However, the data for 2007 was released and approved for the study. The survey is a cross–sectional study with several outcomes on demographic characteristics, household consumption, health, education, housing, social welfare & related programmes, and issues pertaining to persons of prime working age in Jamaica. The survey is conducted on a yearly basis.

3.2.1 The Setting of the Study and Design

The Jamaica Survey of Living Conditions, 2007 (JSLC 2007) was conducted in April 2007 and focused mainly on households. The distribution of the households by parish in the JSLC2007 is presented in Table 3.2. The population distribution in Table 3.2 revealed that there are more people residing in the parish of St. Andrew and St. Catherine. The parishes have more households than other parishes and this is the reason for the higher selection of the enumeration districts in these parishes in Table 2.2.

(Source:http://statinja.gov.jm/Demo_SocialStats/populationbyparish.aspx)

Table 3.2: JSLC 2007 Distribution of Surveyed Household and population by Parish for the Census 2001

Parish	**Number of EDs (PSUs)**			**Number of Households**			**Population**
	Urban	**Rural**	**Total**	**Urban**	**Rural**	**Total**	
Kingston	6	0	6	65	0	65	666,041
St. Andrew	34	4	38	339	41	380	
St. Thomas	2	4	6	22	51	73	94,410
Portland	2	4	6	24	46	70	82,183
St. Mary	1	5	6	11	63	74	114,227
St. Ann	6	4	10	49	65	114	173,232
St. James	6	3	9	62	40	102	184,662
Trelawny	5	3	8	53	36	89	75,558
Hanover	0	6	6	0	73	73	69,874
Westmoreland	2	6	8	21	85	106	144,874
St. Elizabeth	2	8	10	17	94	111	150,993
Manchester	5	7	12	58	93	151	190,812
Clarendon	4	10	14	53	129	182	246,322
St. Catherine	22	9	31	285	119	404	518,345
Total	**97**	**73**	**170**	**1059**	**935**	**1994**	**2,711,476**

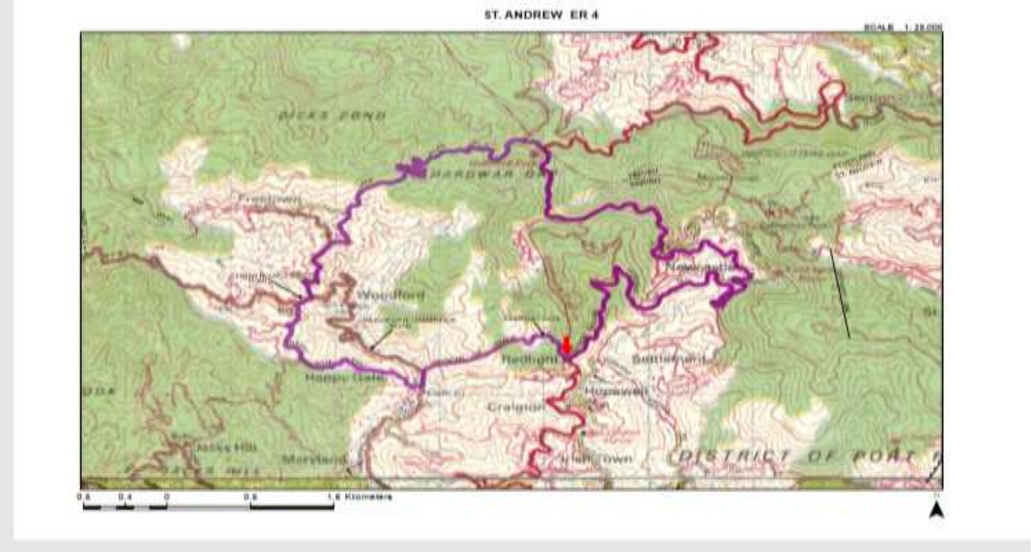
3.2.2 Selection of Enumeration Districts

The JSLC 2007 survey is a complex survey that was designed in two stages. In the first stage, Primary Sampling Units (PSU) are selected from a list as illustrated in Table 3.2. In Jamaica, the Enumeration Districts (ED) are the Primary Sampling Units. The ED is an independent geographic unit sharing common boundaries with a contiguous ED. Except for the parish of Kingston, that has all the EDs being classified as **urban,** every other parish is comprised of both rural and urban EDs. The minimum number of dwellings for EDs in rural areas is 100, while those in urban areas is 150. The second stage of the survey involves the selection of the dwellings using a systematic sampling method with a fixed interval.

3.2.3 JSLC2007 Data Collection

The survey utilizes a face-face to interviewing technique in the data collection. Each interviewer is assigned a specific number of EDs to survey. A map of each ED is provided to aid the interviewer's location of the selected dwelling. (For example, the red boundary in Figure 3.2 illustrates a typical ED map used in the JSLCS and others conducted island wide by STATIN). The STATIN always assigned a team of supervisors to oversee the interviewers on a regular basis.

Figure 3.3: Sample Enumeration District Map - St. Andrew East Rural



3.2.4 JSLC Questionnaire

The JSLC Questionnaire consists of ten modules inclusive of the roaster. The modules are: Health (*A*), Education (*B*), Children items (*C*), Programme for Advancement Through Education and Health, PATH, (*D*), Daily Expenses (*E*), Food Expenses (*F*), Consumption Expenditure (*G*), Non-Consumption Expenditure (*H*), Housing and Related Expenses (*I*).

The food components are found in modules *E*, *F* and *G*. In module *E*, the head of the household was asked if they had bought coal, kerosene oil, and wood. Module *E* also includes questions regarding meals away from home such as meat, poultry, fish meals bought away from home, sandwiches, burgers, milk, other milk-based product, breakfast beverages, fruits, juices, vegetables and drinks in box, bottle or other packaging, and other form of meals such as soups, fresh or frozen beef, pork, fish, chicken purchase and vegetarian meals. Module *F* consists of items of food that was received as gift. The non-food items are found in module *G*. These include personal care items, laundry items and kitchen supplies. The socio-demographic questions on household members and head are located in the roaster module of the questionnaire. Each household is assigned a serial number and this number is applied to all the members of the household.

3.2.5 JSCL2007 Data Source

The data for this study was obtained from the archives of the Sir Arthur Lewis Institute of Social and Economic Studies (SALISES), the University of the West Indies, Mona. SALISES is the archive for Survey data of JSLC Surveys conducted by the Statistical Institute of Jamaica (STATIN) for Academic Research in the West Indies. The permission for the use of the data was approved on behalf of the STATIN by the relevant officers at SALISES. The study data for the JSLC 2007 was extracted from two sources. The first source was the Annual Expenditure data of 1994 households. The second source was the roaster file containing the demographic information of 6,613 household members. The Annual Expenditure data consists of the overall household expenditure on food consumption and other household expenditures. In addition, this Annual Expenditure is also a major component in the baseline analysis for determining the poverty status of a household in Jamaica by the Planning Institute of Jamaica (PIOJ).

3.3 Introduction of the JSLC 2007 Study Variables and Context

Study variables such as the household annual expenditure database include the annual household expenditure of meals purchased away from home (t meal), the size of the household, the constituency, the enumeration district, the parish and the serial number for the household were extracted from the JSLC 2007 Annual Expenditure file. The serial number for each household was used to select each household head and demographic information about the head of the household from the 6,613 household members. The selected demographic variables include of the household head include age, level of education, occupation, employment status and sex.

Tables 3.4 and 3.5 contains the list of the variables, the sample and relevant percentages for each category in the variable. In Table 3.4, of the 1994 households in the sample, 82.6% or 1648 cases provided data on the annual expenditure of meals purchased away from home suggesting that 17.4% cases were missing. Consequently, there are more females than males among the household heads and varying age groups with more middle age individuals in the sample. The household heads are engaged in nine occupational groups. The explorative analysis in the Tables 3.4 and 3.5, also shows that the household heads are employed but the majority did not provide their level of education.

Table 3.4: List of study variables from the JSLC2007

Variables	**n**	**Percentage**
Annual Expenditure of Meal Purchased Away from Home	1648	82.6
Age		
15 – 19	16	0.8
20 – 24	78	3.9
25 – 29	136	6.8
30 – 34	202	10.1
35 – 39	233	11.7
40 – 44	248	12.4
45 – 49	227	11.4
50 – 54	159	8.0
55 – 59	154	7.7
60 – 64	120	6.0
65 – 69	136	6.8
70 – 74	103	5.2
75 – 79	88	4.4
80 – 84	60	3.0
85+	34	1.7
Total	**1994**	**100**
Household Size		
1	451	22.6
2	374	18.8
3	352	17.7
4	269	13.5
5	229	11.5
6	142	7.1
7+	177	8.9
Total	**1994**	**100**
Gender		
Males	1070	53.7
Females	924	46.3
Total	**1994**	**100**

Table 3.5: List of study variables from the JSLC2007

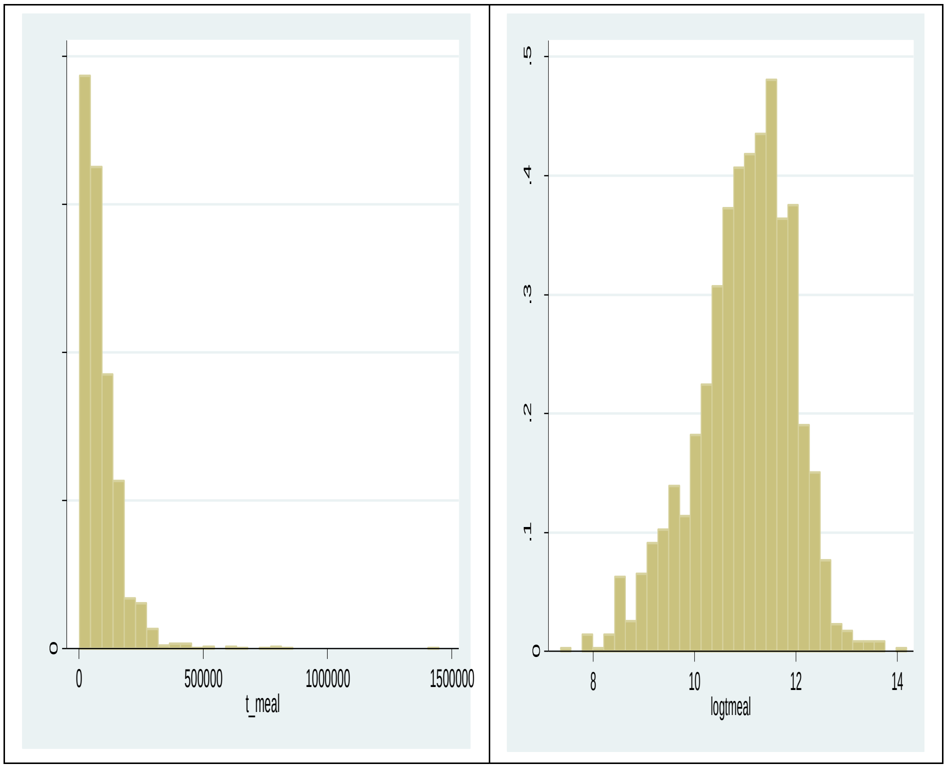
Level of Education	**n**	**Percentage**
High School Level (CXC, CAPE and GCE)	201	10.1
University Level (Degree and other)	120	6.0
None and not stated	1673	83.9
Total	**1994**	**100**
Employment Status		
Employed	1683	84.4
Unemployed	152	7.62
Outside Employment	159	7.97
Total	**1994**	**100**
Occupation		
Skilled Agricultural and Fishery Worker	364	18.25
Non-classified	315	15.8
Craft and Related Trade Workers	276	13.84
Elementary Occupations	275	13.79
Service workers, Shop and Market Sales Workers	267	13.39
Professionals	108	5.42
Plant & Machine Operators & Assemblers and Elementary Occupations	106	5.32
Technicians and Associate Professionals	97	4.86
Legislators, Senior Officials, Managers and Professionals	94	4.71
Clerks	92	4.61
Total	**1994**	**100**

3.3.1 Outcome Variable – Annual Expenditure of Meal Purchased Away from Home

In the JSLC 2007, a total of 1994 household records were sampled. Of the 1994 records only 1,648 had reported the annual expenditure of meal purchased away from home, whilst 346 cases expenditure were missing for different reasons. The highest expenditure of meals purchased away from home in the 1,648 cases was J$1,449,571 with three persons in the household. It is noteworthy that the lowest expenditure reported was J$1564.29 for a household with one person.

In order to understand the spread of the expenditure, Figure 3.3 illustrates the before and after the log of the expenditure was taken as part of the exploratory analysis.

Figure 3.4: Annual expenditure of meal purchased away from home



3.3.2 Exploring the relationships between the covariates and the response variable

The descriptive statistics in Tables 3.6 and 3.7 for the reported expenditure based on the categories in the covariates suggest that some of these variables may have the potential to predict the response variable. For example, a significant difference (F=3.629, p=0.00) was found for the amount of expenditure for the age group which the head of the household belongs. This is illustrated in the box plot (Figure 3.3.1). The household sizes also influence the expenditure as illustrated by the box plot in Figure 3.3.2. The differences in the expenditure due to the household size was found to be significant (F=19.186, p=0.00); Unlike the household size, the gender of the household heads did not influence the household expenditure of meal purchased away from home (t=0.441, p=0.66). This is evident in Figure 3.3.3. The level of education was recoded into “educated” and “not stated” to effectively observe the effect of education on the household expenditure of the meal away from home. Figure 3.3.4 shows that there is significance difference (F=21.38, p=0.00) in the expenditure between the household that has an educated head and household which the head did not state level of education was merged with not stated because this category will not have an effect on the analysis. The three employment categories were recoded to two categories “employed” and “unemployed” to sufficiently understand its influence on the expenditure of meal purchased away from home. Employed household heads had higher expenditure than the unemployed as illustrated Figure 3.3.5 and the test statistic (t=8.270, p=0.00). The fact that some of the occupation categories are closely related and statistical power in the future modeling will be enhanced, the categories were collapsed to three. The first new category is the “Non-Office Related Work (NORW)” consisting of plant and machine operators and assemblers. Household heads who did not state the occupation remains in the “Not Classified” category while those in the Office Related Work were grouped into the third category. The Office related work (ORW) consist of Legislators, Senior Officials and Managers; Clerks; and Professionals. The influence of the recoded occupation vary was found to be significant (F=29.93, p=0.00) and illustrated in Figure 3.3.6.

Table 3.6: Descriptive statistics of the expenditure by the categories in the covariates (Age and Household Size)

Covariate	Sample (n)	Mean	SD
Age			
15 – 19	16	10.92	0.81
20 – 24	78	10.99	0.97
25 – 29	136	11.12	0.98
30 – 34	202	11.17	0.82
35 – 39	233	11.13	0.92
40 – 44	248	11.03	0.91
45 – 49	227	11.01	0.99
50 – 54	159	11.02	0.92
55 – 59	154	10.94	0.96
60 – 64	120	10.76	0.96
65 – 69	136	10.75	1.04
70 – 74	103	10.54	1.04
75 – 79	88	10.69	1.06
80 – 84	60	10.72	1.03
85+	34	11.11	0.60
Total	**1994**		
Household Size			
1	316	10.64	1.10
2	270	10.80	0.95
3	294	10.92	0.92
4	254	11.10	0.86
5	216	11.16	0.84
6	135	11.32	0.84
7+	163	11.38	0.80
Total	**1648**	10.98	0.96

Table 3.7: Descriptive statistics of the expenditure by the categories in the Covariates (Gender, Level of Education, Employment Status and Occupation)

Gender	**N**	**Mean**	**SD**
Males	1070	10.99	0.96
Females	924	10.97	0.96
Total	**1994**		
Level of Education			
High School Level (CXC, CAPE and GCE)	201	10.93	0.97
University Level (Degree and other)	120	11.46	0.83
None and not stated	1673	11.41	0.62
Total	**1994**		
Employment Status			
Employed	1683	11.05	0.93
Unemployed	152	10.52	1.04
Outside Employment	159	10.39	0.98
Total	**1994**		
Occupation Categories			
Skilled Agricultural and Fishery Worker	364	10.42	0.98
Non-classified	315	11.14	0.87
Craft and Related Trade Workers	276	11.17	0.83
Elementary Occupations	275	11.52	0.84
Service workers, Shop and Market Sales Workers	267	10.88	0.92
Professionals	108	11.19	0.86
Plant & Machine Operators & Assemblers and Elementary Occupations	106	11.36	0.73
Technicians and Associate Professionals	97	11.33	0.82
Legislators, Senior Officials, Managers and Professionals	94	11.35	0.86
Clerks	92	10.50	0.97
Total	**1994**		

Figure 3.5: Annual expenditure of meal purchased away from home by age group of the household head

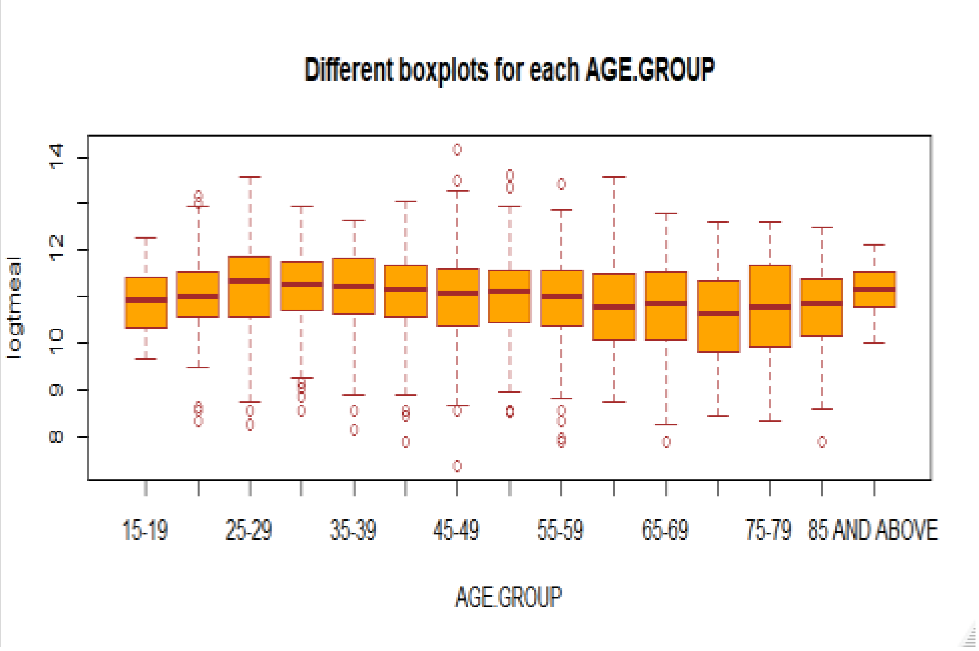


Figure 3.6: Annual expenditure of meal purchased away from home by the household size

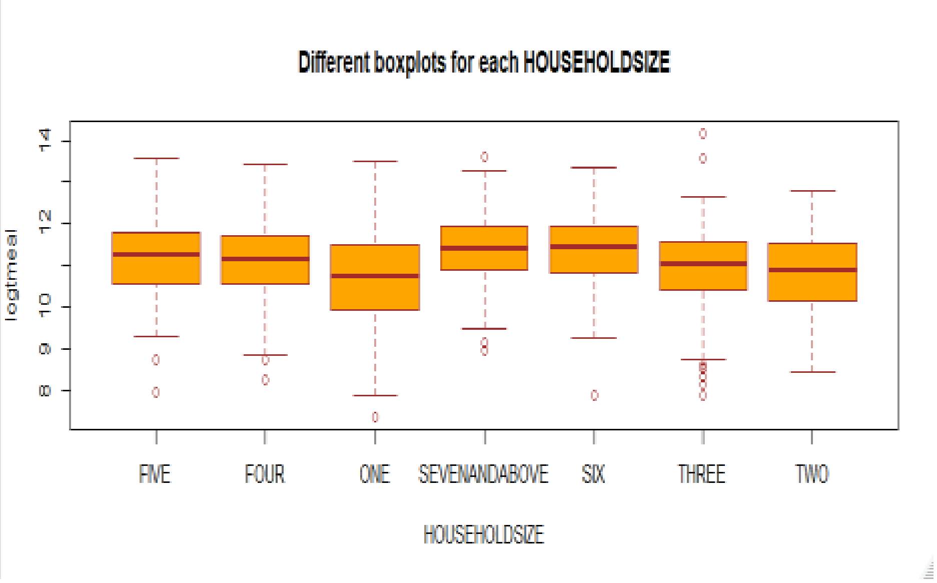


Figure 3.7: Annual expenditure of meal purchased away from home by the sex of the household head

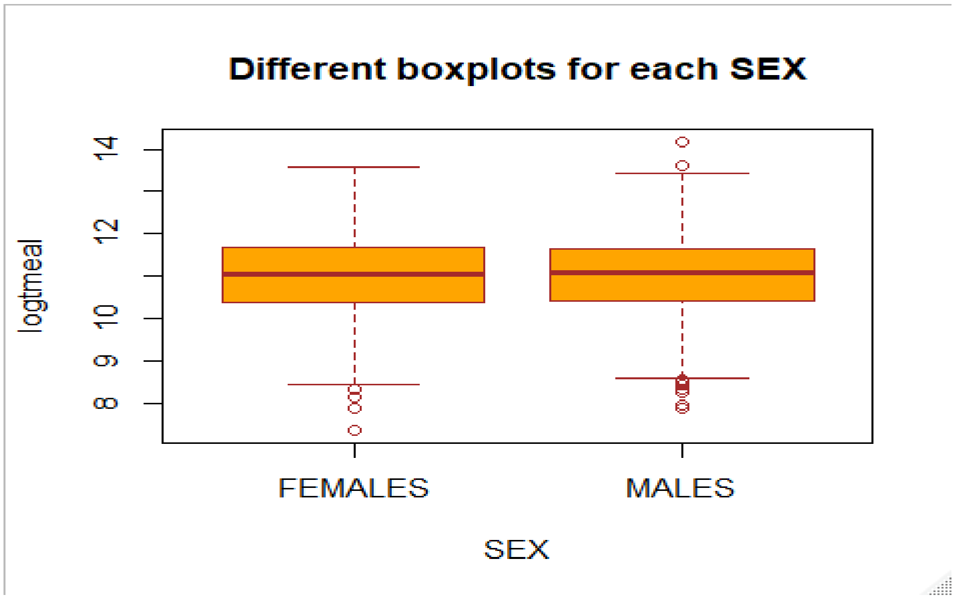


Figure 3.8: Annual expenditure of meal purchased away from home by the education of the household head

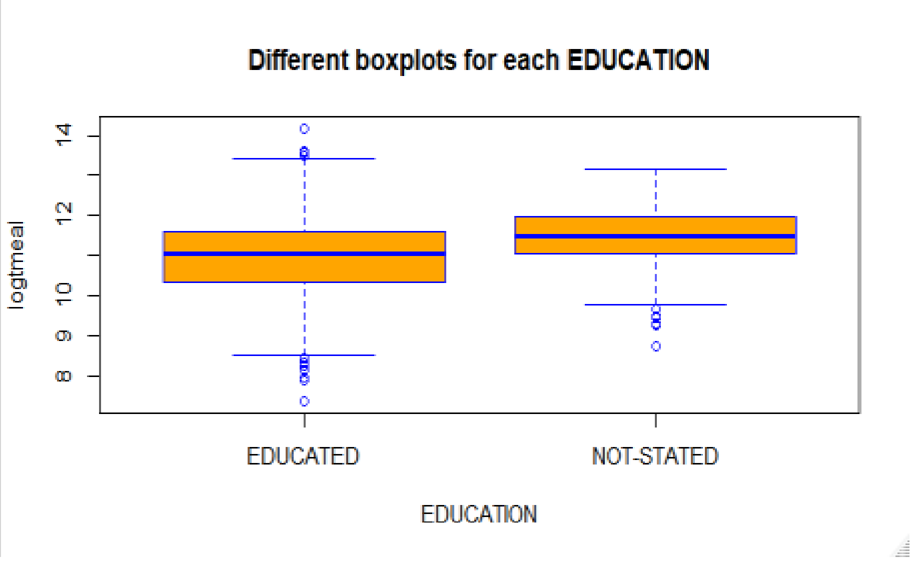


Figure 3.9: Annual expenditure of meal purchased away from home by the employment status of the household head

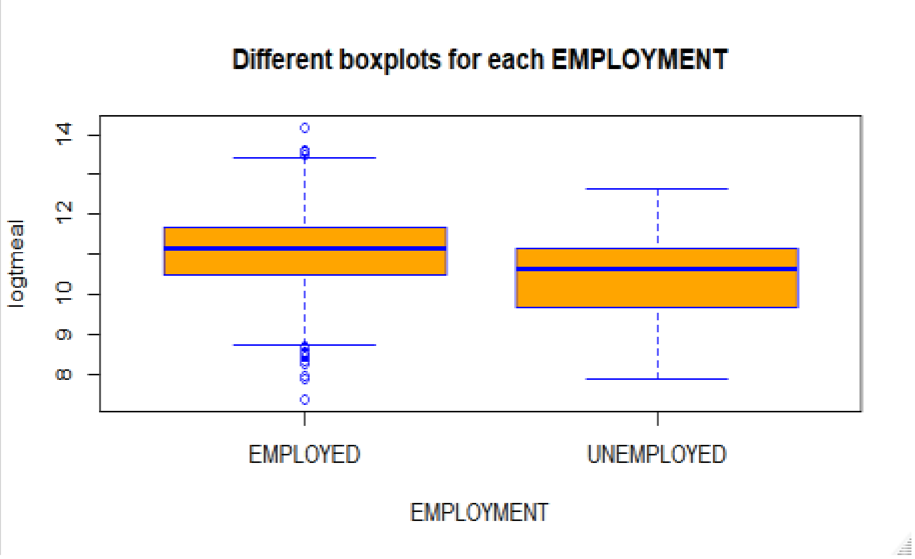
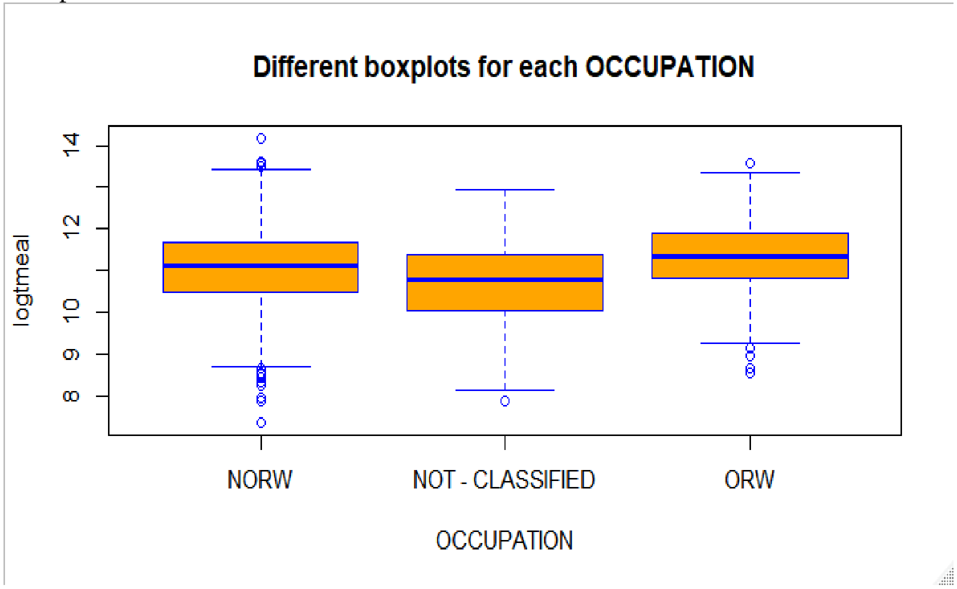


Figure 3.10: Annual expenditure of meal purchased away from home by the occupation of the household head

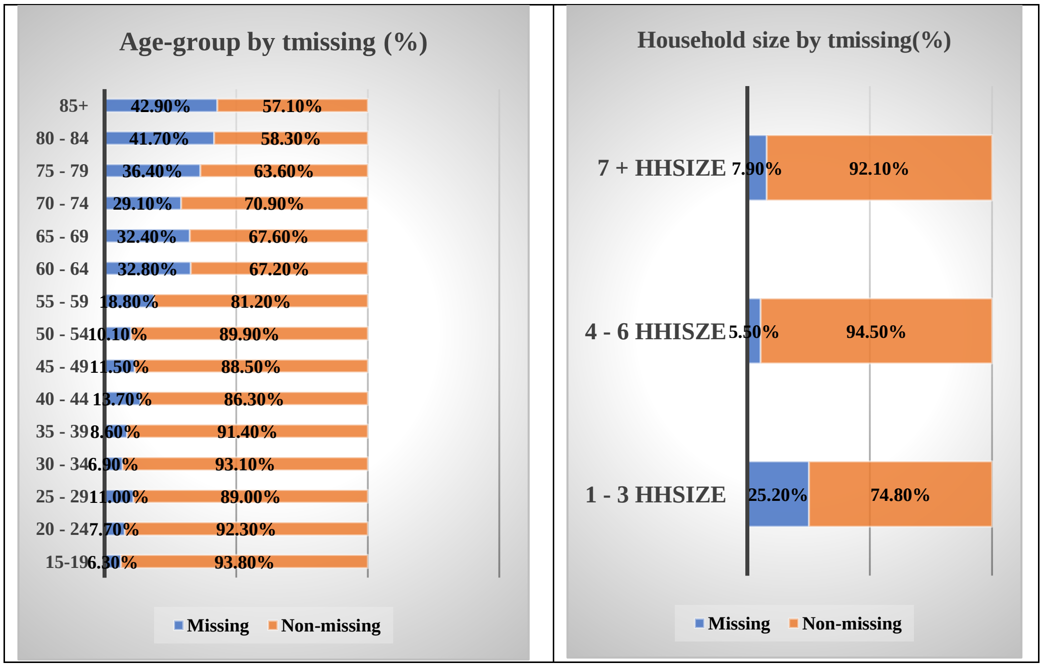


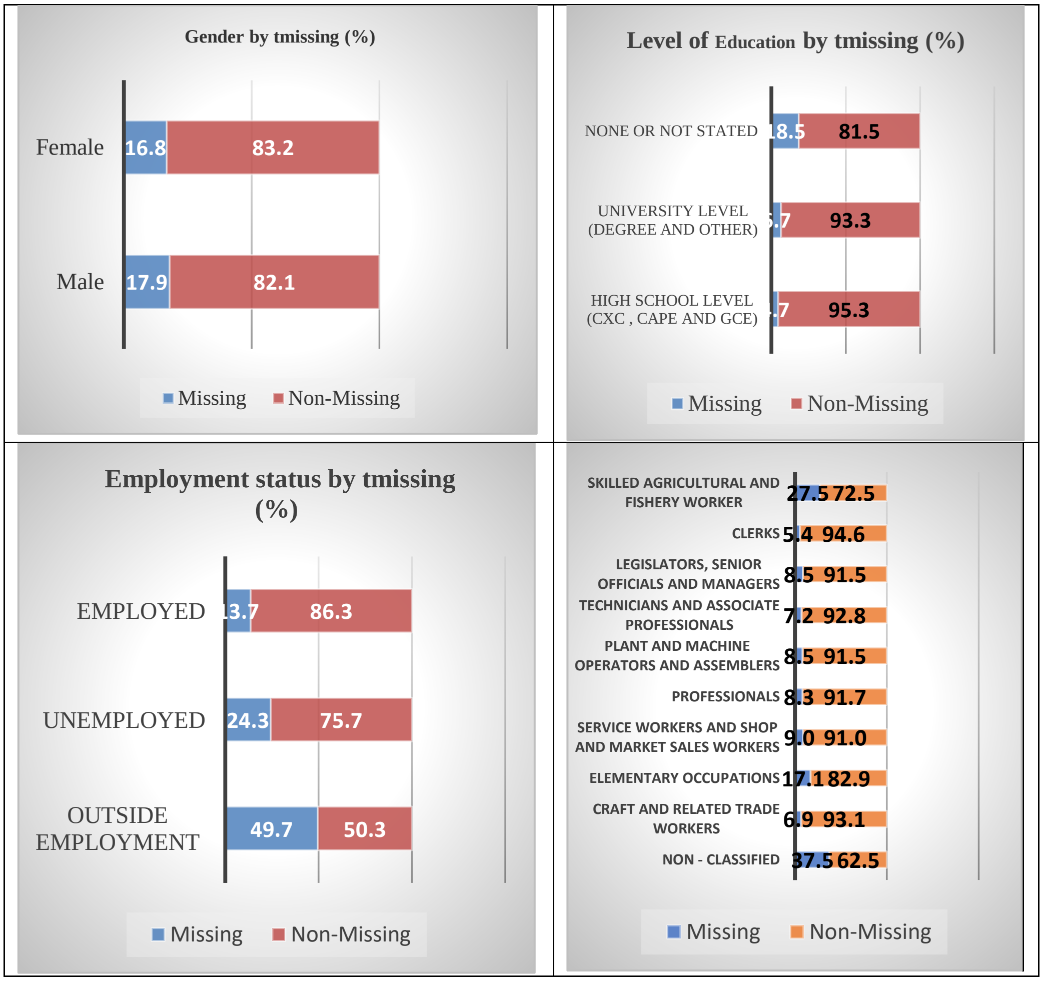
3.3.3 Exploring the missingness in response variable due to the covariates

In this section, the focus is on the missingness in the response variable due to each covariate using Figure 3.4 to illustrates the spread of the missing data in the response variable based on the covariates. The spread of the missingness among the age group revealed that as the age group increases the missingness in the response variable also increases especially as the household head approaches age of retirement (65 – 69 years and above) when compared to the household that is headed by 55-60yrs and less. This suggest that the age covariate has the potential to be a predictor for a missing variable

The spread of missingness among the categories in gender variable is almost the same for male and female headed households while the spread in the other variables as illustrated in Figure 3.8 were not as defined in the pattern as compared with the age covariate.

Figure3.11: Pattern of missingness by predictor variables





3. 4 China Health and Nutrition Survey (CHNS)

In this study, the goal is to investigate the effects of weighting approaches on parameter estimates from multilevel model with emphasis on missing data in the continuous outcome variable. The China Health and Nutrition Survey data have some unique features that facilitates this type of analysis and will serves as the secondary analysis to apply the results to a secondary data. The unique features are: (1) it is a publicly available dataset that scholars from across the world can access; (2) the survey used a multistage, random cluster design in the selection of a stratified probability sample (Zhang et. al, 2014, Popkin et.al, 2009).

3. 4.1 CHNS 1989 and 2011 Selected Variables

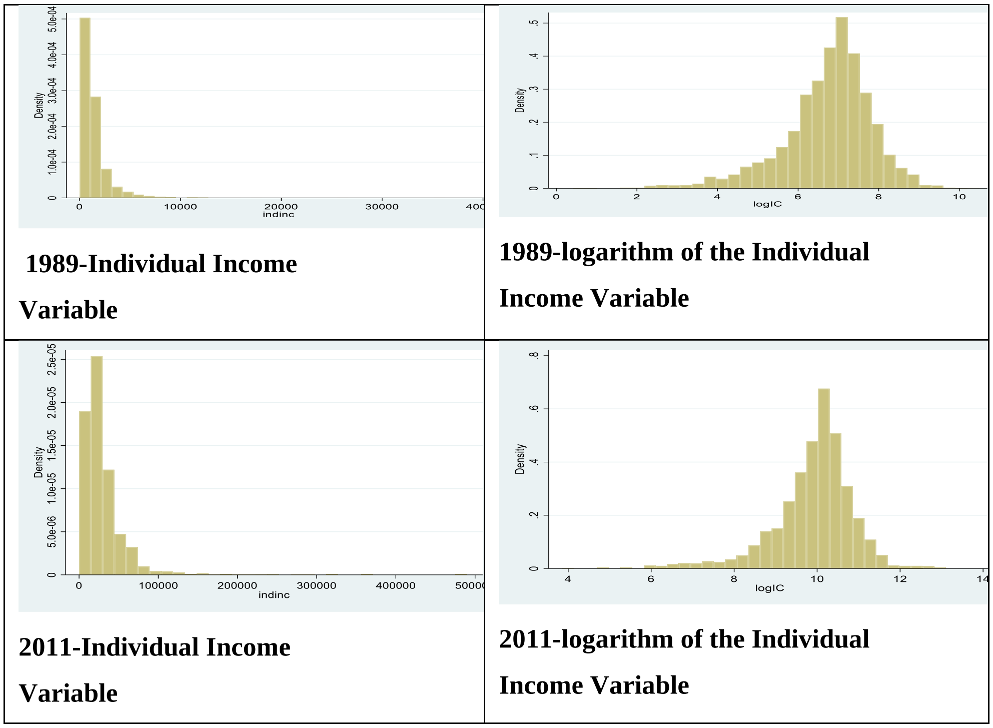
For the purpose of comparison, two different years were selected for analysis, the year 1989 and 2011 datasets were selected with the focus on the dynamics on the reported income of the individuals in the survey as the response variable. In addition, there are two variables on reported income of the individuals in the 1989 and 2011 dataset that were very useful in the analysis (reported income with imputation and without imputation). The survey covers provinces such as Liaoning, Jiangsu, Shandong, Henan, Hubei, Hunan, Guangxi, Guizhou, while in 2011, additional provinces were added such as Beijing, Heilongjiang, Shanghai and Chongqing in 2011 survey. The distribution of the response variables before and after the logarithm of the income variables is illustrated in Figure 3.5 demonstrating the attempts to meet the normal distribution assumption.

In both CHNS1989 and 2011, a total of six similar variables were selected in each data set. The variables consist of reported income with missing data, reported income without missing data, age, gender, marital status, and years of schooling. In the 1989 dataset, 3,052 individuals in the age range of 16yrs to 64yrs were selected from the sample compared with 2,109 sample for 2011 as stated in Table 3.8.

Table 3.8: Selected covariates and response variables from the CHNS 1989 and 2011

	1989		**2011**	
Variable	**n**	**Percentage**	**n**	**Percentage**
Gender				
Male	1606	52.6	984	46.7
Female	1446	47.4	1125	53.3
Total	3052	100	2109	100
Marital status				
Married	2662	87.2	1955	92.7
Not married	390	12.8	154	7.3
Total	3052	100	2109	100
	Mean	**SD**	**Mean**	**SD**
Age	35.03	10.14	45.95	11.94
Years of schooling	15.31	8.81	24.01	7.6
Income without imputation	1371.93	1652.96	28733.88	32007
Income with imputation	1376.01	1655.09	29383.51	32431.9

Figure 3.12: Distribution of the response variable before and after transformation using logarithm



1989-Individual Income Variable

1989-logarithm of the Individual Income Variable

2011-Individual Income Variable

2011-logarithm of the Individual Income Variable

3. 4.2 Exploring effects of the covariates on the response variable -CHNS

The descriptive statistics in Table 3.9 revealed differences in the reported income of the individuals for the selected covariates. The years of schooling has considerable differences in the income of the individuals with 16 years or more than the individuals with less years of education. This affirms the widely view about years of education and income. The income spread by among the age group categories also support the years of schooling effect on the income variable. The high-income earners are in the age group of 30 to 44 years with the expectation that these individuals will have 16 years of more of schooling. The fact that some differences exist in the reported income due to the effect of the covariates, it is expected that these variables are potential predictor in the follow up analysis in Chapter 6.

Table 3.9: Descriptive Statistics for the individual income without imputation for each category in the covariate.

1989				2011		
Variable	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**
Gender						
Males	1606	6.78	1.11	984	10.05	0.96
Females	1446	6.67	1.13	1125	9.78	0.97
Marital status						
Married	2662	6.74	1.11	1955	9.90	0.98
Not married	390	6.62	1.20	154	9.91	0.91
Age group						
15-19	146	6.13	1.10	7	7.85	1.10
20 - 24	321	6.45	1.16	66	9.77	1.22
25 - 29	510	6.68	1.14	171	10.06	0.99
30 - 34	590	6.84	1.09	195	10.16	0.90
35 - 39	520	6.90	1.06	226	10.05	1.05
40 - 44	409	6.91	1.07	280	10.01	0.96
45 - 49	265	6.79	1.05	239	9.94	1.01
50 - 54	169	6.70	1.04	273	9.90	0.79
55 - 59	84	6.41	1.49	359	9.77	0.91
60 - 64	38	6.49	1.30	257	9.66	0.99
65 -69				36	9.64	0.94
Years of schooling						
11-15	840	6.68	1.19	216	9.12	1.00
16+	1599	6.82	1.08	1816	10.04	0.90

Chapter 4

Statistical Methodology

Chapter Outline

This Chapter provides a broad understanding of the Statistical Methodology used in this study. The Statistical Methodology includes concepts such as the theory of parameter estimates for unweighted and weighted linear mixed-effects models, weighting adjustments and associated theory to correct for unequal probability of sample selection and missing data bias in estimates. To fully appreciate these concepts, the first section starts with an elaborate development of multilevel models frame work which can be extended to any level. The second section explains how weights are to be incorporated in the estimation of parameters for weighted multilevel models. The Chapter finishes with a summary section on how the developed weights and methods of parameter estimates would be utilized in Chapters 5 and 6, respectively.

4.1 Multilevel Model

One of the motivations for this research, is the investigation of the effects of weighting adjustments on parameter estimates on linear models consisting of fixed and random effects of continuous outcome with missing data. Despite the fact that multilevel data analysis continues to evolve with novel research for example the work of Bryan and Jenkins (2016), West et.al (2015), Steele and Durrant (2011), Farrell and Ludwig (2008), Zhou et.al (2007), there are more opportunity for continued research in this field, especially the effect of different weighting adjustments on estimates.

a simple hierarchical model to illustrate the principles with further generalization to any hierarchy in the data structure.

In the JSLC2007, J denote the Parishes as the **highest Unit** for $j = 1, 2, 3, ...\mathrm{m}$ and within Parish j, k^{th} denote the Enumeration District (ED) from $1........k_j$. For the CHNS 1989 & 2011 data structure, the notation is repeated by denoting J Provinces as the highest Unit. Within Province j, similarly k^{th} denote the County /City by (j,k) within the province and within each Council /City l^{th} denote the Community by (j,k,l), the lowest Unit.

In the proposed multilevel model, there are n observations split into J Parishes and within j^{th} Parish, there are n_j observations labelled $k = 1, 2, ...n_j$. The response and covariate variables associated with the observations are denoted as y_{jk} and x_{jk} respectively from k^{th} ED within j^{th} Parish . The random errors for the j^{th} Parish and k^{th} ED are represented by $u_j \sim N(0, \sigma_u^2)$ and $\varepsilon_{jk} \sim N(0, \sigma_\varepsilon^2)$ respectively. The assumption is that the data satisfy the model below:

$$y_{jk} = x_{jk}\beta + u_j + \varepsilon_{jk} \tag{4.1}$$

The model in 3.1 is considered as a hierarchical model and may also be refer to a mixed-effect model because of the fixed and random components. The model may be extended to three or more groups in the application to the analysis of the JSLC2007 and CHNS 1989 & 2011.

.

4.2 Weighting Adjustments for the Correction of Sample Selection

To understand the effects of the weights, this section seeks to elaborate on the weighting adjustments that will be used in the analysis to address sampling selection and missing data issues which affects the parameter estimates in the model. In this section, the theory of weight adjustment to correct the unequal probability in the selection of a defined **Unit** and sub populations within the **Unit** are discussed. The Unit selection weight is the first weight to be developed in this section. A Unit is an independent group or cluster in the hierarchy of the sampling process of non-uniform sampling technique. For example, a school and a classroom could be considered as examples of independent Units in the survey of students from a non-uniform sampling. method. Another example is the Parish and Enumeration District in the JSLC 2007 survey. Below are the steps involved in deriving the probability of selecting a Unit and corresponding weight in an independent single stage sampling adapted from Little and Rubin (2002).

In a single stage sampling scheme when n_j units are selected from group j. Denoting the total number of units in group j as N_j , the probability of this selection may be expressed as:

$$\pi_j = n_j / N_j \qquad (4.2)$$

Thus, the weight w_j to address the unequal selection of the Unit is defined as the inverse of the probability (π_j^{-1}).

$$w_j = \pi_j^{-1} = N_j / \mathrm{n}_j \qquad (4.3)$$

However, in a two-stage sampling design where further sampling can take place within a Unit, another probability and corresponding weight will also be required to correct for unequal selection within this Unit. The Valliant, Dever & Kreuter (2013) approach is presented in the section below:

In the first stage of independent sampling , there are N Unit, the total observation , which is split into N_j Unit in group j and N_{jk} Unit in k^{th}

subgroup within group j. Assuming that in the first stage is the same with the previous selection with similar weight as stated in equation (4.3), while in the second stage, n_{jk} units are selected from N_{jk} units with the probability of the second stage of selection defined as:

$$\pi_{k|j} = n_{jk} / N_{jk} \qquad (4.4)$$

The weight to address the unequal selection is the inverse of the probability of the second selection and define as:

$$w_{k|j} = \pi_{k|j}^{-1} = N_{jk} / \mathrm{n}_{jk} \qquad (4.5)$$

Hence, the overall weight to address the first and second selection in the sampling process may be expressed as the product of the inverse probabilities from each stage of the selection as follows:

$$w_{jk} = \pi_j^{-1}\pi_{k|j}^{-1} = N_j N_{jk} / \mathrm{n}_j \, n_{jk} \qquad (4.6)$$

The two - stage independent sampling developed above is synonymous to the selection of n_j schools from a total of N_j schools in the first stage of a selection while in the second stage, n_{jk} classrooms are selected from N_{jk} classrooms.

4.3 Logistic Regression Model and Item Non-Response Weight

The logistic regression model has been recommended for modelling binary response outcomes (McCullagh & Nelder, 1989, pg 110). In this study, the logistic regression model will be used to model the response and non-response taking into account the covariates. Below is an illustration of how the parameters will be estimated.

Let Z_{jki} = 1 if the i^{th} individual with the Enumeration District k did not respond. Suppose that the probability of non-response depends on a covariate x_{jki} with

$$p_{jki} = P(Z_{jki}\, i = 1 \big| X = x_{jki}) \qquad (4.7)$$

Then the logistic regression model relating the outcome variable and the predictor variable can be expressed as:

$$Log_e\left(\frac{p_{jki}}{1-p_{jki}}\right) = \phi_0 + \phi_1 x_{jki} \qquad (4.8)$$

Further to the estimation of the model parameters ϕ_0 , ϕ_1 the probability is thus expressed as:

$$p_{jki} = \frac{e^{\phi_0 + \phi_1 x_{jki}}}{1 + e^{\phi_0 + \phi_1 x_{jki}}} \qquad (4.9)$$

Thus, item non-response weight **(IR)** for each individual i from group k within group j

$$w_{jki} = 1/\,\hat{p}_{jki} \qquad (4.10)$$

4.4 Parameter Estimation

In this Section, parameter estimation will be via maximum likelihood estimation. I start with a simple linear regression model and build up to maximum likelihood estimation for multilevel models and the use of weights in the likelihood process.

4.4.1 Parameter Estimation of Simple Linear Regression Model

Consider a linear relationship between a pair of response y_i and covariate x_i variables for i^{th} observations from a simple random setting which satisfy the regression model below:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon \qquad (4.11)$$

where β_0, β_1 represent the intercept and slope of the model respectively. The random components ε_i are i.i.d with $N(0,\sigma^2)$ and this leads to the y_i independent with the distribution as given. The random error may be expressed for the i^{th} observation as:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \qquad (4.12)$$

The likelihood function of a given observation $\left(x_i, y_i\right)$ and for unknown parameter $\theta = \left(\beta_0, \beta_1, \sigma^2\right)$ based on the model in (4.11) is:

$$L(\theta|y_i) = \prod_i^n f(y_i;\theta) \qquad (4.13)$$

$$L\left(\beta_0, \beta_1, \sigma^2 \,|y_i\right) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}{2\sigma^2}} \qquad (4.14)$$

Then the log-likelihood is

$$\log_e L(\theta|y_i) = \sum_{i=1}^n \log f(y_i;\theta) \qquad (4.15)$$

$$\log_e L = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \left(\beta_0 - \beta_1 x_i\right)\right)^2 \qquad (3.16)$$

By taking the partial derivative of the log -likelihood, a set of equations are generated and set to zero and to find the maximum likelihood estimates of β_0 , β_1 and σ^2 as follows:

$$\frac{\partial \ln L(\mathrm{y}_i;\beta_0,\beta_1,\sigma^2)}{\partial \beta_0} = -\frac{1}{\sigma^2}\sum_i^n \left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \qquad (4.17)$$

$$\frac{\partial \ln L(\mathrm{y}_i;\beta_0,\beta_1,\sigma^2)}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum_i^n \left(y_i - \beta_0 - \beta_1 x_i\right) x_i = 0 \qquad (4.18)$$

$$\frac{\partial \ln L(\mathrm{y}_i;\beta_0,\beta_1,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2 = 0 \quad (4.19)$$

This gives the maximum likelihood estimates (MLEs):

$$\widehat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \left(X^{'}X\right)^{-1} X^{'}Y \qquad (4.20)$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) \qquad (4.21)$$

In some instances, there are situations where the observation errors are independent but not identically distributed that is they do not meet the assumption that $\varepsilon_i \sim N(0,\sigma^2)$, a required assumption for regression models. Therefore, the introduction of a weight ($w_i = 1/\sigma_i^2$), the reciprocal of the observation's error term variance may become necessary to address this problem (Neter, Wasserman and Kutner, 1985, pg.169). This is the basis for the Weighted Least Squares (WLS) Regression Model.

The estimation of the parameters when the weights are introduced is illustrated in the scenario below.

Consider w_i as the diagonal matrix $W(n \times n)$ of weights required for each observation.

$$W = \begin{bmatrix} w_1 & & & & & 0 \\ & w_2 & & & & \\ & & w_3 & & & \\ & & & . & & \\ & & & & . & \\ 0 & & & & & w_n \end{bmatrix} \qquad (4.22)$$

Initially, the likelihood without the weight was (3.13)

$$L(\theta|y_i) = \prod_i^n f(y_i;\theta)$$

With the introduction of the weight the likelihood becomes:

$$L(\theta|y_i) = \prod_i^n f(y_i;\theta)^{w_i} \tag{4.23}$$

$$\log_e L_w\left(\beta_{0w},\beta_{1w},\sigma_w^2|y_i\right) = \sum_{i=1}^n w_i \log f\left(y_i;\theta\right) \tag{4.24}$$

$$\log_e L_w\left(\beta_{0w},\beta_{1w},\sigma_w^2|y_i\right) = \sum_{i=1}^n w_i \log_e\left[\sqrt{\frac{w_i}{2\pi}}\exp^{-\left(\frac{\left(y_i-(\beta_o-\beta_1 x)\right)^2}{2\sigma_w^2}\right)}\right] \tag{4.25}$$

$$\log_e L_w = \frac{nw_i}{2}\log_e w_i - \frac{nw_i}{2}\log 2\pi - \frac{1}{2\sigma_w^2}\sum_{i=1}^n w_i\left(y_i-\left(\beta_0-\beta_1 x_i\right)\right)^2 \tag{4.26}$$

The partial derivative of the log -likelihood are generated as follows:

$$\frac{\partial \ln L(\mathrm{y}_i;\beta_{0w},\beta_{1w},\sigma_w^2)}{\partial\beta_{0w}} = -\frac{1}{\sigma_w^2}\sum_i^n w_i\left(y_i-\beta_{0w}-\beta_{1w}x_i\right)=0 \tag{4.27}$$

$$\frac{\partial \ln L(\mathrm{y}_i;\beta_{0w},\beta_{1w},\sigma_w^2)}{\partial\beta_{1w}} = -\frac{1}{\sigma_w^2}\sum_i^n w_i\left(y_i-\beta_{0w}-\beta_{1w}x_i\right)x_i=0 \tag{4.28}$$

$$\frac{\partial \ln L(\mathrm{y}_i;\beta_{0w},\beta_{1w},\sigma_w^2)}{\partial\sigma_w^2} = -\frac{n}{2\sigma_w^2}+\frac{1}{2\sigma_w^4}\sum_i^n w_i\left(y_i-\beta_{0w}-\beta_{1w}x_i\right)^2=0 \tag{4.29}$$

The Weighted Least Squares regression parameter estimators for $w_i \neq 1$

are:

$$\widehat{\beta}_w = \begin{pmatrix} \beta_{0w} \\ \beta_{1w} \end{pmatrix} = \left(X^{'}WX \right)^{-1} X^{'}WY \tag{4.30}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(Y - X\hat{\beta}\right)' W \left(Y - X\hat{\beta}\right) \tag{4.31}$$

4.3.2 Parameter Estimation of Multilevel Model

The parameter estimation is illustrated with a hierarchical model involving the Parish and the Enumeration District (ED) via the likelihood approach. In the JSLC 2007, there are fourteen Parishes denoted by $j = 1........m$ and also, there are n_j EDs within parish j . The EDs are denoted by $k = 1.........\text{n}_j$ .In the development of a model y_{jk} represents the continuous response from k^{th} Enumeration District within the j^{th} Parish and x_{jk} , the covariate for every y_{jk} observation. The random errors for the j^{th} Parish and k^{th} ED as $u_j \sim N(0,\sigma_u^2)$ and $\varepsilon_{jk} \sim N(0,\sigma_\varepsilon^2)$ respectively for the purpose of developing a mixed-effect model that fits the structure of the data. The proposed model is:

$$y_{jk} = x_{jk}\beta + u_j + \varepsilon_{jk} \tag{4.32}$$

For the model in (4.32), the likelihood will be expressed as a joint distribution for the observed data in the Parishes and EDs with two sources of randomness since the values of u and y are known as follows:

$$f\left(y \middle| u,\beta,\sigma^2,\sigma_u^2\right) = \prod_{j=1}^{m}\prod_{k=1}^{n_j} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left\{y_{jk} - x_{jk}\beta - u_j\right\}^2\right) \tag{4.33}$$

and

$$f\,(\text{u}\left|\beta,\sigma^2,\sigma_u^2\right.) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_U}} \exp\left(-\frac{1}{2\sigma_U^2}u_j^2\right) \tag{4.34}$$

By combining 3.34 and 3.35, the likelihood becomes:

$$=\int \prod_{j=1}^{m}\frac{1}{\sqrt{2\pi\sigma_U}}\exp\left(-\frac{1}{2\sigma_U^2}u_j^2\right)\prod_{k=1}^{n_j}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2\sigma^2}\left\{y_{jk}-x_{jk}\beta-u_j\right\}^2\right)du \qquad (4.35)$$

$$=\prod_{j=1}^{m}\left\{\int \frac{1}{\sqrt{2\pi\sigma_U}}\exp\left(-\frac{1}{2\sigma_U^2}u_j^2\right)\prod_{k=1}^{n_j}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2\sigma^2}\left\{y_{jk}-x_{jk}\beta-u_j\right\}^2\right)du_j\right\} \qquad (4.36)$$

To further simplify the likelihood, the unknown parameters are denoted as $\theta=\left(\beta,\delta,\delta_U\right)$, $\delta=1/\sigma^2$, $\delta_U=1/\sigma_U^2$, for $N=\sum_{j=1}^{m}n_j$

By substituting in 3.36 the revised likelihood given **u is**:

$$L(\theta;\mathrm{y},\mathrm{u})=\prod_{j=1}^{m}\frac{\sqrt{\delta_U}}{\sqrt{2\pi}}\exp\left(-\frac{\delta_U}{2}u_j^2\right)\prod_{k=1}^{n_j}\frac{\sqrt{\delta}}{\sqrt{2\pi}}\exp\left(-\frac{\delta}{2}\left\{y_{jk}-x_{jk}\beta-u_j\right\}^2\right) \quad (4.37)$$

Then the log-likelihood will be:

$$l(\theta;y,u)=\mathrm{K}+\frac{m}{2}\log\delta_U-\frac{\delta_U}{2}\sum_{j=1}^{m}u_j^2+\frac{N}{2}\log\delta-\frac{\delta}{2}\sum_{j=1}^{m}\sum_{k=1}^{n_j}\left\{y_{jk}-x_{jk}\beta-u_j\right\}^2 \quad (4.38)$$

If $K=-\frac{1}{2}\log(2\pi)\left(m+N\right)$, the log-likelihood further reduces to:

$$=\mathrm{K}+\frac{m}{2}\log\delta_U-\frac{\delta_U}{2}\sum_{j=1}^{m}u_j^2+\frac{N}{2}\log\delta-\frac{\delta}{2}\sum_{j=1}^{m}\sum_{k=1}^{n_j}\left\{\left(y_{jk}-x_{jk}\beta\right)^2-u_j(y_{jk}-x_{jk}\beta)+u_j^2\right\} \quad (4.39)$$

This is now applied in the Expectation - Maximization algorithm to find Maximum Likelihood for the unknown parameters $\theta=\left(\beta,\delta,\delta_U\right)$. The first step in the process is the identification of the distribution of u_j as:

$$f(u_j\left|\theta,y,u_{-j}\right.)\propto\exp\left(-\frac{\delta_U}{2}u_j^2\right)\exp\left(-\frac{\delta}{2}\sum_{k=1}^{n_j}\left\{y_{jk}-x_{jk}\beta-u_j\right\}^2\right) \qquad (4.40)$$

$$\propto\exp\left(-\frac{\delta_U}{2}u_j^2-\frac{\delta n_j}{2}u_j^2+\delta u_j\sum_{k=1}^{n_j}\left(y_{jk}-x_{jk}\beta\right)\right) \qquad (4.41)$$

$$= \exp\left(-\frac{1}{2}\left\{ u_j^2\left(\delta_U + \delta n_j\right) - 2u_j \sum_{k=1}^{n_j}\left(y_{jk} - x_{jk}\beta\right) \right\} \right) \qquad (4.42)$$

Hence, distribution thus can be conveniently expressed as follows:

$$U_j \left| \theta, y, u_{-j} \right. \sim N\left(\sum_{k=1}^{n_j}\left(y_{jk} - x_{jk}\beta\right) / \left(\delta_U + \delta n_j\right), 1/\left(\delta_U + \delta n_j\right) \right) \quad (4.43)$$

In order to find the values that maximizes the log-likelihood, the E-step in the EM -algorithm was started with initial value of $u_j = 0$ for all j in (3.38). This is followed by the computation of $E\left[u_j \left| \theta, y, u_{-j} \right. \right]$ and $E\left[u_j^2 \left| \theta, y, u_{-j} \right. \right]$ in place of u_j and u_j^2 in (3.38) for the M-step. This is repeated for the re-computation of the E-step and M-step until convergence is achieved.

4.3.4 Pseudo Maximum Likelihood Estimation Method

The JSLC2007 scenario is used to illustrate the calculation of the weight in the parameter estimation while addressing unequal sample selection and for non-response. The Pseudo Maximum Likelihood has been proposed as the estimator for the weighted parameter (See Rozi, 2013; Skinner., 1998; Binder, 1983). This procedure will be used in this study. One of the issues in the non-uniform sampling is the unequal sampling or non-response. The following section is an illustration on the use of the weight in the correction of the unequal sampling of the k^{th} ED within the j^{th} Parish. Consider $w_{k|j}$ as the weight for the correction of unequal sampling of the k^{th} ED within the j^{th} Parish. The weight is defined as

$w_{k|j} = \pi_{k|j}^{-1} = N_{jk} / \mathrm{n}_{jk}$. Using the earlier parameterisation $\theta = \left(\beta, \delta, \delta_U\right)$,

This can be expressed as the pseudolikelihood:

$$L_P\left(\theta; \mathrm{y}, \mathrm{u}\right) = \prod_{j=1}^{m} \prod_{k}^{n_j} f\,(\mathrm{y}, \mathrm{u}; \theta)^{w_{k|j}} \qquad (4.45)$$

Recalling the solution for the log-likelihood in 3.39, similar derivation will be express for the log-pseudolikelihood in 3.46 below

$$l_p(\theta; y,u) = \mathrm{K} + \frac{m}{2} w_{k|j} \log \delta_U - \frac{\delta_U}{2} \sum_{j=1}^{m} w_{k|j} u_j^2 + \frac{N}{2} w_{k|j} \log \delta - \frac{\delta}{2} \sum_{j=1}^{m} \sum_{k=1}^{n_j} w_{k|j} \left\{ y_{jk} - x_{jk}\beta - u_j \right\}^2$$

If $w_{k|j} = 1$, the log-pseudolikelihood will equal the log-likelihood.

Hence, the parameter estimates that will maximize the log-pseudolikelihood, are identified from the EM algorithm explained earlier in this section.

4.4 Summary of the Statistical Methodology

In this Chapter, the weighting adjustments and associated parameter estimators for unweighted and weighted multilevel model are elaborated and will be utilized in the simulation study and the analyses of JSLC2007 and CHNS 1989 &2011.In Chapters 5 and 6.

Chapter 5

Simulation study

Chapter Outline

In this chapter, I present a simulation study to assess the various weighting methods in obtaining a reliable parameter estimate. Existing techniques were used, along with a variety of choices to estimate parameters, including a range of statistical assessment tools to assess the reliability of the parameter estimate. The chapter is concluded with a discussion and recommendations for addressing non-response and sample selection bias.

5.1 Rationale and Purpose of the simulation study

The rationale for the simulation study is the assessment of the weighting method to address non-response and sampling bias problems in a survey data. The use of simulation studies to evaluate statistical models and methods is a common practice (see e.g., Asparauhov, 2006; Burton et al., 2006; Chiou & Muller, 2005; Yucel et al., 2018). This simulation experiment is designed for the primary purpose of providing answers to the following questions:

(a) What is the performance of weighting methods for sample selection and non-response problems?

(b) Under what condition(s) does a weighting method perform best?

5.2 Specific objectives of the simulation study

The specific objectives of this simulation investigation are to:

(a) Simulate the properties of a complex survey data using a multilevel linear model with a continuous outcome so that the best performed weighting methods may be identified.

(b) Evaluate the performance of the weighting methods by using the parameter estimates from different scenarios in the simulation study.

(c) Identify similarities and dissimilarities in the parameter estimates from weighting methods for different missing mechanisms and different proportions of omissions to identify the conditions for the weighting adjustment method.

5.3 Review of the Survey Weights in the Simulation Study

In the simulation study, four weights will be evaluated. This will include, Enumeration District (ED) selection weight (w_j); Household selection weight from the ED ($w_{k|j}$) within a Parish, design weight (w_{jk}), the item non-response weight (w_{jki}). These weighting adjustments are illustrated in the following sections to complement the review.

5.3.1 Illustration of the weights using a case from the JSLC 2007

For the purpose of illustration, consider an observation i of a household with a household head of age 71 selected from an ED within the Parish. There are 217 households in the ED of which 16 households were sampled but 14 households completed the survey instrument from a Parish in the island. In this Parish, there are 126 EDs of which 4 EDs were selected.

The summary of the data is provided as follows:

$N_j = 126$ is the total number of EDs in Parish j .

$n_j = 4$ is the number of EDs sampled in Parish j .

Hence, the probability of selecting a given ED in Parish j is calculated as:

$$\pi_j = \frac{n_j}{N_j} = \frac{4}{126} = 0.031746 \quad .$$

$$w_j^{PS} = p_j^{-1} = \frac{N_j}{n_j} = 31.5 \, .$$

The **31.5** represents the weights for every ED sampled in that Parish.

To calculate the household weight, the procedure is as follows:

Let $N_{jk} = 217$ be the total number of the households or dwellings within k^{th} ED within the j^{th} Parish.

Let $n_{jk} = 14$ be the number of households that responded in k^{th} ED within the j^{th} Parish.

$$\pi_{k|j} = \frac{n_{jk}}{N_{jk}} = \frac{14}{217} = 0.064516$$

Thus, the weight for the household and subsequently household head selection in the ED within a Parish:

$$w_{k|j}^{HS} = \pi_{k|j}^{-1} = \frac{N_{jk}}{n_{jk}} = 15.5 \, .$$

The combined weights for ED and Household selections are denoted as (**PHS)** in the simulation study and calculated as:

$$w_{jk}^{PHS} = \pi_j^{-1}\pi_{k|j}^{-1} = \frac{N_j N_{jk}}{n_j n_{jk}} = \frac{(126)(217)}{(4)(14)} = 488.25 .$$

In the calculation of the item non-response **(IR)**, the age of the subject, 71 was applied in the logistic model below to deduce the propensity of responding as:

$$\hat{p}_{jki} = \hat{p}_{jki}(71\ ) = \frac{\exp(\mathrm{X}'_{jki}\ \beta)}{1+\exp(X'_{jki}\beta)} = 0.6905963 \text{for an individual.}$$

The above probability was calculated for the household heads in the k^{th} *ED* for household *l=2,3,4,5,6,8,9, 10, 11, 12,13,14* in Table 5.0.

Table 5.0: Summary of household propensity to respond

Household i	age_{jkl}	p_{jkl}
1	*71*	*0.69060*
2	*41*	*0.88502*
3	*71*	*0.6906*
4	*39*	*0.89316*
5	*62*	*0.76392*
6	*59*	*0.78552*
7	*Nil*	*Nil*
8	*41*	*0.88502*
9	*66*	*0.73287*
10	*34*	*0.91131*
11	*Nil*	*Nil*
12	*67*	*0.72472*
13	*45*	*0.86713*
14	*44*	*0.87181*
15	*41*	*0.88502*
16	*40*	*0.88916*

Thus, the response probability in for the households in k^{th} ED within the j^{th} Parish is:

$$\overline{p}_{avg_{jk}} = 0.8197043 .$$

The item non response weight **(IR)** is thus calculated as follows:

$$w_{jki}^{IR} = \left(\overline{p}_{avg_{jki}}\right)^{-1} = 1.219952 .$$

The proposed weighting methods is summarized in Table 5.1

Table 5.1: Sampling and non-response weights

Weight	**Symbol**
Sampling Selection Weights	
ED selection weight	PS
Household selection weight	HS
ED and Household selection weights	PHS
Non-Response Weight	
Item non-response weight	IR
Sampling Selection Weight Adjusted For Missing Data	SWAM

5.3.2 Scaled weights

The investigation on weighting scale in multilevel modelling continues to generate a debate in the literature (see Carle, 2009; Rabe–Hesketh & Skrondal, 2006). These studies focus on binary responses while investigating the effects of weight scaling in multilevel modelling. Rabe–Hesketh and Skrondal (2006) concluded that "for small clusters, estimated random -intercept variance was found to be biased" while Carle (2009, pg.6) also concurred that "scaled weighted estimates and standard errors differed slightly from unweighted analyses and observed that the differences were minimal and did not lead to different inferential conclusions". These conclusions were not available for multilevel models with a continuous outcome.

The findings from these studies serve as the rationale for the inclusion of the effect of weight scaling in multilevel model using a continuous outcome. Accordingly, scaling methods proposed by Carle (2009, pg.9) were adopted for this study.

Scaling methods were applied to the design weights (**PHS**), item non-response weight (**IR)** and **SWAM.** The scaling methods are described as follows:

5.3.2.1 Method A

For method A, w_{ij}^{P} for *p= PHS, IR & SWAM* is denoted as the original weight to be scaled while $w_{ij}^{P.\mathrm{A}}$ is denoted as the scaled weight for a given weight. In this method the weights are scaled to the size of n_j , that is, the number of sampled units in cluster j . Specifically, the formula below is used to calculate scaled weight A for i^{th} household in the k^{th} ED within j^{th} Parish.

$$w_{jkl}^{P.A} = w_{jkl}^{P}\left(\frac{n_j}{\sum_i w_{jkl}^{P}}\right) \tag{5.1}$$

5.3.2.2 Method B

In the scaling process for Method B, the scale weight is denoted as $w_{ij}^{P.\mathrm{B}}$. When the original weight of w_{ij}^{P} was scaled to the sum of the weights in cluster j , the results are indicated in the formula below:

$$w_{jkl}^{P.B} = w_{jkl}^{P}\left(\frac{\sum_i w_{jkl}^{P}}{\sum_i \left(w_{jkl}^{P}\right)^2}\right) \tag{5.2}$$

5.4 Simulation Plan

Four simulation scenarios were considered. Each held a combination of 20%, 40% and 60% missing data with missing mechanism MAR and MNAR. The model below will be used to simulate the data using parameters of a prior analysis of the JSLC2007 as a guide:

$$y_{jki} = x_{jki}^{T}\beta + u_{j} + v_{jk} + \varepsilon_{jki} \tag{5.3}$$

In the first scenario, the chosen parameter for the model was $\beta' = (\beta_0 = 6,\ \beta_1 = -0.080)$ when age is the independent variable, while $x_{jki}^{T} = (1, age_{jki})$ of the individuals in the survey. For the second scenario, the parameter changes to $\alpha' = (\alpha_0 = 6, \alpha_1 = 5)$when the sex is the independent variable, and $x_{jki}^{T} = (1, sex_{jki})$. The third scenario model parameters combination of the age and sex covariates while the fourth is the extension of the third model with the inclusion of household size, education, employment and occupation of the individuals in the surveys. In the four models, the random components where $u_j \sim N(0,3^2)$, $v_{jk} \sim N(0,2^2)$ and $\varepsilon_{jki} \sim N(0,1^2)$ for parish *j*, ED, *k* and subject *jkl* levels respectively. The simulated data generating mechanism were represented as: $Y_{1,\mathrm{jki}}$, $Y_{2,\mathrm{jki}}$, $Y_{3,\mathrm{jki}}$,$Y_{4,\mathrm{jki}}$ for each model. These simulated data sets were used in the recovery of the weighted parameter estimates for reliability of the weighting methods.

5.5 Missing at Random (MAR) and Missing Not at Random (MNAR)

The missing at random (**MAR)** mechanism was generated using a binomial function *Miss,* where$Miss \sim Bin(n, p)$, for *n* =1994 and *p =0.2, 0.4, 0.6.* The logistic model *loMiss* was then applied to create the missing not at random (**MNAR)** using the R-Software Version 3.5.2

The logistic regression model function used to create the scenario for **MNAR** is:

$$loMiss = \frac{e^{(\Delta_0 + \Delta_1 * age)}}{1 + e^{(\Delta_0 + \Delta_1 * age)}} \tag{5.4}$$

The model parameters were set at different values for Δ_0 = 0.07, 0.88,1.91 and Δ_1 = - 0.03; while aiming for 20%, 40% and 60% missing respectively, based on the data generation model in the simulation study. These parameters were carefully chosen and varied so as to achieve the rate of missing in relation to the population of interest. A total of N=100 repetitions was simulated for each weighted multilevel model containing 1994 cases in each scenario where MNAR and MAR represented rate of missing (20%, 40%, and 60%).

5.6 Simulation Statistical Assessment Tools

To evaluate the bias in the parameter estimates, the indicator for performance of the weighting adjustment methods, the error in the estimates were calculated to identify the relative closeness of the estimated and true parameters. The calculation was expanded by estimating the average errors to determine the root mean square error (RMSE) for each weighting method. Yucel et al.'s (2018) percent bias (PB) approach assessed the performance of estimated parameters relative to the true value. The strength of this approach was the availability of large PB value for large error size and vice versa for small error size. Additionally, in order for a method to be effective, the PB should not exceed 5 percent (Demirtas, Freels, & Yucel, 2008).

The RMSE and PB were derived for each weighting methods as follows:

Let T be the true value of the parameter in the data generation models and $E(\hat{T})$ is the average estimated for parameter of interest (PI) from the evaluative models.

The average is estimated thus:

$$E(\hat{T}) = \bar{\hat{T}} = \frac{1}{N}\sum_{i=1}^{N=100}\hat{T} \qquad (5.5)$$

for i=1………N=100 .

Ultimately leading to the percent bias defined as:

$$PB = 100 \times \left|\left(E(\hat{T}) - T\right)/T\right| \qquad (5.6)$$

For the RMSE, $\hat{T}_i$ is defined as the estimated parameter from each model and T_i be the true parameter, the errors were calculated as

$$\mathrm{e}_i = T_i - \hat{T}_i \qquad (5.7)$$

Hence, the mean square error MSE = $\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{e}_{\mathrm{i}}^2$ (5.8)

for i=1,2,3,4…………N =100 .

And finally RMSE = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}e_i^2}$ (5.9)

5.7 Results of the Simulation Study

5.7.1 Simulation Study without Incorporation of Varying Intraclass Coefficient

There are series of results originating from the parameter estimates for the fixed and random components in the simulation experiment study. The result of the parameter estimates for the continuous and categorical variable were used as a generalization for all continuous and categorical variables. The recovered parameter estimates for age and sex variables were selected as the illustration of the results. To understand the findings on the recovery of the parameter based on the weighting methods, Table 5.2, an excerpt was created from more detailed findings in Tables 5.5 and 5.6. This illustration is from the series of simulation on parameter estimates for continuous variable using the data generating mechanism Y_{1jkl} and a true value $\beta_1 = -0.08$.

Table 5.2.: Root mean square error (RMSE) for the ED selection weight (PS), Household selection weight (HS), Combination of ED and Household selection weight (PHS) and sampling weight adjusted for item non response (SWAM).

	MAR			**MNAR**		
	RMSE VALUES			**RMSE VALUES**		
WM	**20%**	**40%**	**60%**	**20%**	**40%**	**60%**
PS	0.0721	0.0722	0.0720	0.0719	0.0720	0.0721
HS	0.0723	0.0722	0.0722	0.0720	0.0722	0.0718
PHS	0.0723	0.0724	0.0722	0.0719	0.0721	0.0720
IR	0.0722	0.0722	0.0720	0.0719	0.0722	0.0719
SWAM	**0.0019**	**0.0736**	**0.0313**	**0.0017**	**0.0022**	**0.0028**
Unweighted	**0.0721**	**0.0722**	**0.0720**	**0.0719**	**0.0721**	**0.0719**

According to Table 5.2, the root mean square error for the PS, HS, PHS and IR weighting methods were relatively similar to the unweighted method. But this changes when the sampling weights is adjusted for missing data using the item nonresponse weight (SWAM), the root mean square error were relatively smaller compared with the other weighing methods without adjustment for missing data.

In Tables 5.5 and 5.6, the detailed results for MAR and MNAR are presented respectively for the PS, HS, & PHS weighting method. The rate of missing for Y_{3jki} and $Y_{4,\mathrm{jki}}$ were set at 20%, 40% and 60%. For the MAR, the weighted parameter estimates from $\hat{\lambda}_{1_{UP}}$, $\hat{\lambda}_{1_{PS}}$, $\hat{\lambda}_{1_{HS}}$, $\hat{\lambda}_{1_{PHS}}$ and $\hat{\lambda}_{1_{IR}}$ all were within acceptable limit of the percent bias (<5%) while the least root mean square error was observed for item non-response weight **(IR) f**or Y_{3jki} . For the overall results, of the three selection weights (PS, HS, PHS), the **HS** appears to outperform the other selection weights regardless of the scaling methods A or B. All the parameter estimates for age from the weighted model were found to be significant (p<0.005). Consequently, the parameter estimates from the simulation scenario involving the data generating mechanism that contains a categorical covariate variable sex, the parameter estimates from all the different weighting methods $\hat{\lambda}_{2_{UP}}$, $\hat{\lambda}_{2_{PS}}$, $\hat{\lambda}_{2_{HS}}$, $\hat{\lambda}_{2_{PHS}}$, and $\hat{\lambda}_{2_{IR}}$ were also found to be within the acceptable limit of a 5% bias. The result also revealed that **IR** has the best performed weighting method with the least root mean square error. The findings also revealed that the root mean square error increases as the rate of missing increases. Also, these parameter estimates were not significant (p > 0.05) unlike the coefficient for the continuous variable.

In the $Y_{4,\mathrm{jki}}$ data generation model, all the weighting methods were found to be reliable for the continuous covariate because the parameter estimates ($\tau_{1_{UP}}$, $\hat{\tau}_{1_{PS}}$, $\hat{\tau}_{1_{HS}}$, $\hat{\tau}_{1_{PHS}}$, and $\hat{\tau}_{1_{IR}}$) have percent bias within the acceptable limit. This is in contrast to the parameter estimates for the categorical covariate in the model for sex variable estimates ($\tau_{3_{UP}}$, $\hat{\tau}_{3_{PS}}$, $\hat{\tau}_{3_{HS}}$, $\hat{\tau}_{3_{PHS}}$,and $\hat{\tau}_{3_{IR}}$) which shows that the weighting methods were unreliable. In the MNAR scenario, the results were similar as the MAR scenario with item non-response weight (**IR**) was found to have the least root mean square at the 20% rate of missing followed by the household selection weight (**HS**). In the MAR and MNAR scenarios for Y_{1jki} and $Y_{2,\mathrm{jki}}$ data generating mechanism were also found to produced reliable parameter estimates for both continuous and categorical covariates (See Tables 5.5 & 5.6)

Tables 5.7 and 5.8 contains the findings of the scaled weights using Method A, while similar findings for Method B can be found in Appendix E. The results from Method A were similar to the findings from Method B for both MAR and MNAR scenarios. The scaled weights in the simulation study included **PHSA, IRA, PHSB**, **IRB**, which investigated the effects of scaling. In addition, both Tables 5.7 and 5.8, the weights methods were found to be reliable because the root mean square were found to be minimal while the percent bias was within the acceptable limit of less than 5% based on estimates from Y_{1jki}, $Y_{2,\mathrm{jki}}$, Y_{3jki} . In $Y_{4,\mathrm{jki}}$, the weighting methods were unreliable for the estimates of the categorical covariate ($\tau_{3_{UP}}$, $\hat{\tau}_{3_{PHSA}}$ and $\hat{\tau}_{3_{IRA}}$). The item non-response weight (**IRA**) was observed to have the least root mean square error among all the scaled weights followed by the design weight (**PHSA**). Similar results are contained in other tables in Appendix E.

This section contains the results of the performance of the weight with regards to the random components. Three random components were examined; u_j for the parish cluster, v_{jk} for the PSU cluster and ε_{jki} for the household level. The true value for these random components were 3, 2, and 1, respectively. The results for the random components u_j and v_{jk} from the MAR and MNAR scenarios revealed that the parameter estimates were unreliable and the weighting methods performed poorly. These results are not illustrated in this section but can be found in Appendix E. In Table 5.15, the estimates from MAR for ε_{jki} are presented. A thorough analysis of these results in Table 5.15 revealed that all the weights performed and were reliable except for estimates of the weights from the household selection (**HS**). Of all the weighting methods, the **PS** and **IR** weights were found to produce better estimates than other weights in the estimation for ε_{jki} .

5.7.2 Simulation Study with Incorporation of Varying Intraclass Correlation Coefficient

The objective of this simulation is to determine the performance of the weighting adjustment methods under varying intra-correlation coefficient. In this simulation study different values of the random components were selected arbitrarily to deduce different values of the Intra-class correlation coefficients (ICCs) at the Parish, the highest and Enumeration District which is referred to as the Primary Sampling Unit respectively. For this simulation three scenarios were investigated with different values of the random components under Missing at Random (MAR) and Missing Not at Random (MNAR) while proportion of missing cases in the dependent variable was set at 20%. For the first scenario the random components were $u_j \sim N(0,4^2)$, $v_{jk} \sim N(0,7^2)$ and $\varepsilon_{jki} \sim N(0,10^2)$. These values were reversed in the second scenario as $u_j \sim N(0,10^2)$, $v_{jk} \sim N(0,7^2)$ and $\varepsilon_{jki} \sim N(0,4^2)$. In the third scenario, lower values were selected $u_j \sim N(0,5^2)$, $v_{jk} \sim N(0,4^2)$ and $\varepsilon_{jki} \sim N(0,3^2)$. In the earlier simulation study, the random components were $u_j \sim N(0,3^2)$, $v_{jk} \sim N(0,2^2)$ and $\varepsilon_{jki} \sim N(0,1^2)$ for parish *j*, the Primary Sampling Unit (PSU), *k* and subject *jkl* levels respectively. Hence the Intra-class correlation coefficient (ICC) values for the Parish, the Primary Sampling Units (PSU) and individual levels remained constant during the simulation.

Findings

The findings Table 5.3 involving 20% proportion missing at random (MAR) revealed that the root mean square error reduces for the estimates from the four weighting adjustments investigated as the ICC values increases. Similar findings were observed in Table 5.4 with same 20% proportion of missing not at random (MNAR). Of the three scenarios investigated, the findings for the 0.4586 or 45.86% ICC for the parish and 0.7738 or 77.38% for the PSU generated based on the random component for: $u_j \sim N(0,5^2)$, $v_{jk} \sim N(0,4^2)$ and $\varepsilon_{jki} \sim N(0,3^2)$ had the most reduced root mean square errors for each of the estimates investigated. At the 20% proportion of missing for either MAR or MNAR, the findings suggest that the varying ICC values in the weighted multilevel model can affect the performance of the estimator. The incorporation of the ICCs values will be necessary in simulation studies to enable reliable estimates. Further simulation studies with higher proportions of missing under MAR and MNAR mechanisms will be necessary to generalize beyond the 20% missing cases.

Table 5.3: Parameter Estimate of a weighted three level model incorporating different values of ICC under MAR Mechanism at 20% proportion of missing

MISSING AT RANDOM (MAR)								
			ICC -AVERAGE					
U	V	E	ICC- PARISH	ICC- PSU	WM	TV	AE	RMSE
4	7	10	0.0795	0.3671	HS	-0.0800	-0.0800	0.0151
					PS	-0.0800	-0.0803	0.0158
					PHS	-0.0800	-0.0807	0.0173
					IR	-0.0800	-0.0802	0.0139
5	4	3	0.4586	0.7738	HS	-0.0800	-0.0794	0.0056
					PS	-0.0800	-0.0797	0.0053
					PHS	-0.0800	-0.0798	0.0058
					IR	-0.0800	-0.0793	0.0050
10	7	4	0.5578	0.8803	HS	-0.0800	-0.0793	0.0070
					PS	-0.0800	-0.0793	0.0067
					PHS	-0.0800	-0.0789	0.0074
					IR	-0.0800	-0.0793	0.0062

Weighting Adjustment Methods (WM); True Value (TV); Average Estimate (AE); Root Mean Square Error (RMSE); Intraclass correlation (ICC)

Table 5.4: Parameter Estimate of a weighted three level model incorporating different values of ICC under MNAR Mechanism at 20% proportion of missing

MISSING NOT AT RANDOM (MNAR)								
			ICC -AVERAGE					
U	**V**	**E**	**ICC- PARISH**	**ICC- PSU**	**WM**	**TV**	**AE**	**RMSE**
4	7	10	0.0845	0.3777	HS	-0.0800	-0.0797	0.0174
					PS	-0.0800	-0.0786	0.0191
					PHS	-0.0800	-0.0786	0.0196
					IR	-0.0800	-0.0790	0.0162
5	4	3	0.4280	0.7663	HS	-0.0800	-0.0807	0.0055
					PS	-0.0800	-0.0810	0.0053
					PHS	-0.0800	-0.0809	0.0059
					IR	-0.0800	-0.0807	0.0049
10	7	4	0.5499	0.8789	HS	-0.0800	-0.0809	0.0067
					PS	-0.0800	-0.0814	0.0072
					PHS	-0.0800	-0.0810	0.0076
					IR	-0.0800	-0.0811	0.0062

5.8 Summary of the simulation study

The simulation study provided clearer understanding of parameter estimation from weighted multilevel models with varying rate of missing. The findings in relation to the earlier questions which facilitated the design of the simulation study under the different headings are summarized in three broad headings: performance of weighting adjustment; and conditions under which a weighting adjustment will perform best.

5.9.1 Performance of the Weighting Methods for Sample Selection Bias

The findings from the simulation study on the sample selection weights **- PS**, **HS,** and **PHS -** show that the selection weights seldom produced reliable estimates, especially for continuous covariates in any multilevel model. Based on the information provided on the survey design, any of the weights in the study could be used to address sample selection bias. Of the three selection weights used in the study, the **PS** was the first level weight most performed, most reliable and with the lower root mean square error. This was followed by the **HS,** which was **the second level weight** for the majority of the scenarios in the simulation study. However, the design weight, the combination of both PS and HS which addresses unequal selections at both stages of the survey design exhibited higher root mean square error and was found to be unreliable.

5.9.2 Performance of the Weighting Methods for Non-response Problems

The item non-response weight was developed to address non-response bias in the simulation study. The results from the scenarios in the simulation study revealed that, item non-response weight (**IR**) was found to be reliable estimate from the results. This was confirmed from the root mean square errors obtained from the simulation. One of the advantages of the item non-response is that the weight development relies on the available information in the data unlike the unit non-response adjustment and design weights. Both the unit non-response adjustment and design weights will require the use of survey design information which may not be available for secondary analysis.

5.9.3 Performance of the Sampling Weighting Adjusted for Missing Data using the Item Non-response Weight

Adjusting for missing data at the individual level in addition to the adjustment for the bias due to unequal probabilities of the level one and two Units of selection revealed much lower root mean square error for the recovery of the true parameter estimates. Especially under 20% proportion of missing for simulated data with continuous, categorical data generating mechanisms (Tables 5.11 and 5.12) in comparison with the simulated data from multivariate data generating mechanism. Of all the estimates, adjusting for sample selection probabilities and missing data produce better estimates when the relevant information on the survey design is available.

5.9.4 Performance of Scaled Weight

Three weights (**PHS**, **IR & SWAM)** were scaled to the size of the cluster and also to the sum of the weights within the cluster to investigate the effects of scaling. The reliability of the estimated parameters from the scaled weights were found to be similar for the MAR and MNAR scenarios from either method of scaling. The estimates from the scaled item non-response weight were observed to have the least root mean square and judged to be the best of all the methods. This suggests that the scaling added no value to the parameter estimates apart from a reduction in error value.

5.9.5 Identifiable conditions for weighted multilevel models in the presence of missingness/ omission

The analyses of the results of the simulation study show that the root mean square error for the 20% rate of missing cases were relatively reduced when compared with the values for 40% and 60% rate of missing cases. This implies that the fewer the cases of missing, the better the performance of the weighting methodology. It was deduced that at the 20% rate of missing, all the parameter estimates were found to be reliable except for the categorical variable in the weighted multivariable model which was not significant.

5.9.6 Performance of the random components

The simulation experiment affirms poor performance of the weights except for besides the item non-response weight in the parameter estimates of the random components. The item non-response weight produced reliable estimates for the individual level random component. Other random components for the other cluster levels such as the parish and ED were poorly estimated by the weights.

5.9.6 Lessons Learnt from the simulation study

The conclusion from this simulation study is that item non-response weighting appears to be reliable across the scenarios of missing at random and missing not at random scenarios for rates of missing under 20%. Similarly, weighting methods for known design information at the cluster levels can be considered for multilevel modelling of outcome variable to address sample selection bias. The weighting adjustment for the unit non-response appears to be less performed in the simulation when compared with the item non-response and sample selection weights at different levels.

5.9.7 Future simulation study on weighting methods

This simulation study may be extended in future. This would be possible through the simulation of multiple imputation of a given outcome variable with missing data from a survey data at a specified varying rate of missing cases. The objective would be to obtain similar root mean square errors from weighting methods to determine the conditions at which both methods will produce same magnitude of errors in the parameter estimates under the missing at random (MAR) and missing not at random (MNAR) scenarios.

Any future study will seek to find root mean square error at which parameter estimates from either method will be similar thereby enabling equivalence in the handling of missing data in the analysis of survey data. This research will be a useful tool to guide modelling of continuous, binary and multinomial variables in multilevel modelling.

Table 5.5: Summary of the simulation results when the missing mechanism is MAR with missing rates of 20%, 40% and 60% for each variable for sex - age and multivariable covariates models

MISSING AT RANDOM (MAR)												
				20%			40%			60%		
DGM	PI	WM	TV	AE	PB	RMSE	AE	PB	RMSE	AE	PB	RMSE
Y_{3jkl}	λ_1	UP	5.00000	4.99870	0.02600	0.06111	4.98970	0.20600	0.05964	4.99225	0.15500	0.08342
		PS	5.00000	4.99630	0.07400	0.06855	4.99020	0.19600	0.06532	4.98952	0.20960	0.09891
		HS	5.00000	4.99590	0.08200	0.06183	4.98810	0.23800	0.06615	4.99613	0.07740	0.08706
		PHS	5.00000	4.99374	0.12520	0.06933	4.98752	0.24960	0.07092	4.99698	0.06040	0.10127
		IR	5.00000	4.99880	0.02400	0.06145	4.98998	0.20040	0.05981	4.99227	0.15460	0.08302
	λ_2	UP	-0.08000	-0.07982	0.22500	0.00153	-0.08020	0.25000	0.00155	-0.07933	0.83750	0.00250
		PS	-0.08000	-0.07985	0.18750	0.00167	-0.08025	0.31250	0.00181	-0.07946	0.67500	0.00269
		HS	-0.08000	-0.07988	0.15000	0.00177	-0.08031	0.38750	0.00177	-0.07930	0.87500	0.00288
		PHS	-0.08000	-0.07993	0.08750	0.00183	-0.08032	0.40000	0.00199	-0.07943	0.71750	0.00305
		IR	-0.08000	-0.07983	0.20875	0.00155	-0.08020	0.24750	0.00156	-0.07934	0.82375	0.00249
Y_{4jkl}	τ_1	UP	-0.00780	-0.00804	3.10769	0.00195	-0.00791	1.39615	0.00231	-0.00765	1.93282	0.00283
		PS	-0.00780	-0.00798	2.24872	0.00233	-0.00785	0.58654	0.00252	-0.00758	2.80897	0.00310
		HS	-0.00780	-0.00798	2.24872	0.00202	-0.00803	3.00769	0.00261	-0.00769	1.38923	0.00289
		PHS	-0.00780	-0.00797	2.17708	0.00231	-0.00793	1.68000	0.00279	-0.00761	2.49064	0.00336
		IR	-0.00780	-0.00804	3.04872	0.00196	-0.00791	1.41667	0.00231	-0.00765	1.90538	0.00286
	τ_3	UP	-0.02500	-0.03161	**26.44281**	0.04828	-0.02472	1.12492	0.06602	-0.01943	**22.27556**	0.08408
		PS	-0.02500	-0.03102	**24.08832**	0.05546	-0.01979	**20.83184**	0.07338	-0.01891	**24.36880**	0.09345
		HS	-0.02500	-0.03284	**31.34240**	0.05279	-0.02550	1.98520	0.07462	-0.02294	**8.25708**	0.09305
		PHS	-0.02500	-0.03257	**30.26380**	0.06178	-0.02169	**13.24407**	0.08097	-0.02222	**11.13884**	0.10605
		IR	-0.02500	-0.03139	**25.54420**	0.04801	-0.02487	0.53520	0.06589	-0.01876	**24.95794**	0.08414

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.6: Summary of the simulation results when the missing mechanism is MNAR with missing rates of 20%, 40% and 60% for each variable for sex - age and multivariable covariates models

MISSING NOT AT RANDOM (MNAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{3jkl}	λ_1	UP	5.00000	5.01085	0.21700	0.05398	5.00473	0.09460	0.06083	4.99372	0.12560	0.07657
		PS	5.00000	5.00717	0.14340	0.05750	5.00875	0.17500	0.07285	4.99344	0.13120	0.08181
		HS	5.00000	5.00712	0.14240	0.05866	5.00903	0.18060	0.07219	4.99566	0.08680	0.08046
		PHS	5.00000	5.00167	0.03340	0.06491	5.01005	0.20100	0.08286	4.99434	0.11320	0.08652
		IR	5.00000	5.01091	0.21820	0.05387	5.00466	0.09320	0.06089	4.99373	0.12540	0.07689
	λ_2	UP	-0.08000	-0.08004	0.04625	0.00166	-0.07994	0.07000	0.00166	-0.08004	0.04625	0.00240
		PS	-0.08000	-0.07991	0.11500	0.00175	-0.07987	0.15875	0.00193	-0.07991	0.11500	0.00276
		HS	-0.08000	-0.07998	0.02000	0.00192	-0.08009	0.10625	0.00191	-0.07998	0.02000	0.00259
		PHS	-0.08000	-0.07991	0.11375	0.00194	-0.07999	0.01375	0.00207	-0.07991	0.11375	0.00299
		IR	-0.08000	-0.08003	0.03250	0.00166	-0.07995	0.06625	0.00166	-0.08003	0.03250	0.00241
Y_{4jkl}	τ_1	UP	-0.00780	-0.00786	0.74359	0.00188	-0.00809	3.76282	0.00214	-0.00771	1.19436	0.00285
		PS	-0.00780	-0.00782	0.31667	0.00209	-0.00801	2.67949	0.00228	-0.00875	12.14885	0.00306
		HS	-0.00780	-0.00790	1.28974	0.00200	-0.00824	5.60051	0.00258	-0.00877	12.45769	0.00315
		PHS	-0.00780	-0.00783	0.39359	0.00219	-0.00811	3.99038	0.00271	-0.00886	13.54808	0.00319
		IR	-0.00780	-0.00786	0.81410	0.00189	-0.00809	3.70897	0.00214	-0.00891	14.23667	0.00303
	τ_3	UP	-0.02500	-0.02113	**15.46564**	0.05411	-0.03152	**26.09620**	0.00428	-0.01896	**24.17556**	0.09031
		PS	-0.02500	-0.01881	**24.76920**	0.06303	-0.03226	**29.05008**	0.00591	-0.02902	**16.06308**	0.09224
		HS	-0.02500	-0.02060	**17.59700**	0.06193	-0.02977	**19.06336**	0.00528	-0.02678	**7.13454**	0.09143
		PHS	-0.02500	-0.01858	**25.69439**	0.06947	-0.02942	**17.69712**	0.08314	-0.03014	**20.55140**	0.10314
		IR	-0.02500	-0.02122	**15.11280**	0.05396	-0.03124	**24.97204**	0.06514	-0.02654	**6.15692**	0.08750

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.7: Parameter estimates from weighted multilevel models(scaled method A) from the missing mechanism is MAR with missing rates of 20% , 40% and 60% for each data simulated from sex- age and multivariable covariate models

SCALED METHOD A -MISSING AT RANDOM (MAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	β_1	UP	-0.08000	-0.08016	0.20000	0.07217	-0.08018	0.22500	**0.07221**	-0.07997	0.03750	0.07201
		PHSA	-0.08000	-0.08026	0.32875	0.07229	-0.080234	0.29250	0.07227	-0.08020	0.25375	0.07226
		IRA	-0.08000	-0.08019	0.23250	0.07220	-0.08020	0.24625	0.07223	-0.07996	0.04875	0.07201
Y_{2jkl}	α_1	UP	5.00000	5.00522	0.10440	0.05698	4.98987	0.20260	0.07230	4.99605	0.07900	0.07970
		PHSA	5.00000	5.00517	0.10340	0.07482	4.99105	0.17900	0.09008	4.99397	0.12060	0.10908
		IRA	5.00000	5.00380	0.07600	0.05910	4.99119	0.17620	0.07054	4.99376	0.12480	0.08212
Y_{3jkl}	λ_1	UP	5.00000	4.99870	0.02600	0.06111	4.98966	0.20680	0.05964	4.99225	0.15500	0.08342
		PHSA	5.00000	4.99277	0.14460	0.07165	4.98786	0.24280	0.07236	4.99893	0.02140	0.10013
		IRA	5.00000	4.99786	0.04280	0.06402	4.99118	0.17640	0.06046	4.99527	0.09460	0.08373
	λ_2	UP	-0.08000	-0.07982	0.22000	0.00153	-0.08019	0.23750	0.00155	-0.07933	0.84125	0.00250
		PHSA	-0.08000	-0.07994	0.07875	0.00185	-0.08029	0.35750	0.00199	-0.07943	0.71500	0.00303
		IRA	-0.08000	-0.07982	0.22000	0.00159	-0.08021	0.26250	0.00163	-0.07935	0.81500	0.00250
Y_{4jkl}	τ_1	UP	-0.00780	-0.00786	0.74359	0.00193	-0.00791	1.39615	0.00231	-0.00765	1.93282	0.00285
		PHSA	-0.00780	-0.00786	0.71538	0.00230	-0.00793	1.63846	0.00266	-0.00756	3.11679	0.00342
		IRA	-0.00780	-0.00757	2.96538	0.00200	-0.00789	1.11154	0.00229	-0.00762	2.25449	0.00304
	τ_3	UP	-0.02500	-0.03161	**26.44281**	0.04828	-0.02472	1.12492	0.06602	-0.01943	**22.27556**	0.08408
		PHSA	-0.02500	-0.03230	**29.19424**	0.06440	-0.02003	**19.87948**	0.07984	-0.02236	**10.55372**	0.10528
		IRA	-0.02500	-0.03041	**21.62552**	0.04915	-0.02423	**3.07960**	0.06526	-0.02014	**19.44496**	0.08565

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.8: Parameter estimates from weighted multilevel models (scaled method A) from the missing mechanism is MNAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex-age and multivariable covariate models

SCALED METHOD A - MISSING NOT AT RANDOM (MNAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	β_1	UP	-0.08000	-0.07992	0.10000	0.07194	-0.08014	0.17250	0.07217	-0.07989	0.13750	0.07194
		PHSA	-0.08000	-0.07988	0.15125	0.07191	-0.07998	0.02500	0.07201	-0.07993	0.09000	0.07199
		IRA	-0.08000	-0.07999	0.01750	0.07200	-0.08009	0.11125	0.07212	-0.07986	0.17125	0.07192
Y_{2jkl}	α_1	UP	5.00000	4.99988	0.00240	0.05911	4.99467	0.10660	0.06040	5.01574	0.31480	0.09223
		PHSA	5.00000	4.99901	0.01980	0.06597	4.98983	0.20340	0.07269	5.02193	0.43860	0.10538
		IRA	5.00000	5.00129	0.02580	0.05973	4.99612	0.07760	0.06139	5.01647	0.32940	0.09568
Y_{3jkl}	λ_1	UP	5.00000	5.01085	0.21700	0.05398	5.00473	0.09460	0.06083	4.99372	0.12560	0.07657
		PHSA	5.00000	5.00576	0.11520	0.06409	5.00968	0.19360	0.08148	4.99564	0.08720	0.08539
		IRA	5.00000	5.01202	0.24040	0.05306	5.00507	0.10140	0.06302	4.99345	0.13100	0.07459
	λ_2	UP	-0.08000	-0.08034	0.42750	0.00166	-0.07994	0.07000	0.00166	-0.08004	0.04625	0.00240
		PHSA	-0.08000	-0.08042	0.52125	0.00193	-0.07998	0.03125	0.00209	-0.07995	0.06500	0.00295
		IRA	-0.08000	-0.08037	0.45625	0.00170	-0.07994	0.07750	0.00175	-0.08007	0.09250	0.00238
Y_{4jkl}	τ_1	UP	-0.00780	-0.00786	0.74359	0.00188	-0.00809	3.76282	0.00213	-0.00771	1.19436	0.00286
		PHSA	-0.00780	-0.00789	1.12821	0.00214	-0.00809	3.71564	0.00245	-0.00886	13.63577	0.00323
		IRA	-0.00780	-0.00773	0.86795	0.00205	-0.00814	4.37949	0.00210	-0.00870	11.50731	0.00293
	τ_3	UP	-0.02500	-0.02113	**15.46564**	0.05411	-0.03152	**26.09620**	0.06543	-0.01896	**24.17556**	0.09031
		PHSA	-0.02500	-0.01318	**47.28836**	0.07103	-0.02806	**12.22680**	0.08129	-0.03227	**29.07360**	0.10098
		IRA	-0.02500	-0.02150	**14.01160**	0.05368	-0.03050	**22.01076**	0.06624	-0.02963	**18.52600**	0.08170

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; ; PHSA=ED and Household selection weight using scaled Method A; IRA=Item non-response weight using scaled Method A

Table 5.9: Random Component – estimates from weighted multilevel models from the missing mechanism is MAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex- age and multivariable covariate models

MISSING AT RANDOM (MAR)-E												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	**E**	UP	1.00000	0.99779	0.22119	0.01824	0.99976	0.02384	0.02041	0.99734	0.26604	0.03157
		PS	1.00000	0.99806	0.19430	0.02244	0.92759	7.24080	0.02499	0.99458	0.54192	0.03657
		HS	1.00000	0.94483	**5.51740**	0.05830	0.92749	**7.25131**	0.07568	0.88442	**11.55830**	0.11931
		PHS	1.00000	0.99782	0.21833	0.02388	0.99924	0.07562	0.02755	0.99352	0.64814	0.03882
		IR	1.00000	0.99777	0.22319	0.01827	0.99977	0.02340	0.02035	0.99719	0.28091	0.03161
Y_{2jkl}	E	UP	1.00000	0.99538	0.46173	0.02015	0.99873	0.12680	0.02205	1.00159	0.15913	0.02671
		PS	1.00000	0.99757	0.24325	0.02182	1.00042	0.04244	0.02483	1.00357	0.35730	0.02885
		HS	1.00000	0.94207	**5.79348**	0.06154	0.92472	**7.52834**	0.07827	0.89000	**10.99964**	0.11232
		PHS	1.00000	0.99715	0.28461	0.02360	0.99945	0.05516	0.02491	1.00285	0.28543	0.02891
		IR	1.00000	0.99530	0.47006	0.02018	0.99881	0.11906	0.02218	1.00141	0.14107	0.02688
Y_{3jkl}	E	UP	1.00000	0.99818	0.18217	0.01880	0.99955	0.04455	0.02088	0.99361	0.63923	0.02713
		PS	1.00000	0.99792	0.20794	0.01852	0.99797	0.20314	0.02430	0.99286	0.71360	0.03150
		HS	1.00000	0.94509	**5.49118**	0.05854	0.92626	**7.37393**	0.07667	0.88194	**11.80576**	0.12055
		PHS	1.00000	0.99809	0.19088	0.02094	0.96182	3.81821	0.05995	0.99245	0.75469	0.03258
		IR	1.00000	0.99819	0.18059	0.01882	0.99956	0.04385	0.02078	0.99363	0.63691	0.02698
Y_{4jkl}	E	UP	1.00000	0.99818	0.18217	0.01880	0.99955	0.04455	0.02088	0.99361	0.63923	0.02713
		PS	1.00000	0.99792	0.20794	0.01852	0.99797	0.20314	0.02430	0.99286	0.71360	0.03150
		HS	1.00000	0.94509	**5.49118**	0.05854	0.92626	**7.37393**	0.07667	0.88194	**11.80576**	0.12055
		PHS	1.00000	0.99809	0.19088	0.02094	0.96182	3.81821	0.05995	0.99245	0.75469	0.03258
		IR	1.00000	0.99819	0.18059	0.01882	0.99956	0.04385	0.02078	0.99363	0.63691	0.02698

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.10: Parameter estimates from weighted multilevel models using sampling selection weight adjusted for missing data (SWAM) from the Missing At Random missing mechanism with missing rates of 20% , 40% and 60% for data simulated from age -only, sex-only, sex-age, multivariable covariate models

MISSING AT RANDOM (MAR)												
				20%			**40%**			**60%**		
	PI	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	α_1	SWAM	-0.08	-0.07993	0.0837	0.0019	-0.0801	0.05875	0.0736	-0.07984	0.2041	0.0313
Y_{2jkl}	β_1	SWAM	5	4.9953	0.0940	0.0000	4.9885	0.2296	0.0000	5.02054	0.40911	0.1053
Y_{3jkl}	λ_1	SWAM	5	4.9952	0.0954	0.0684	5.0108	0.2168	0.0829	4.9959	0.0806	0.0863
	λ_2	SWAM	-0.08	-0.0804	0.5375	0.0019	-0.0799	0.0125	0.0021	-0.07994	0.0776	0.0030
Y_{4jkl}	τ_1	SWAM	-0.008	-0.0078	1.8350	0.0022	-0.0080	0.1644	0.0026	-0.00790	1.2919	0.0032
	τ_3	SWAM	-0.025	-0.01869	**25.2176**	0.0689	-0.0308	**23.3450**	0.0800	-0.01612	**55.0522**	0.0983

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; SWAM=Sampling Weight Adjusted for Missing Data

Table 5.11: Parameter estimates from weighted multilevel models using sampling weight adjusted for missing data (SWAM) from the Not Missing At Random missing mechanism with missing rates of 20% , 40% and 60% for data simulated from age -only, sex-only, sex-age, multivariable covariate models

MISSING NOT AT RANDOM (MNAR)												
				20%			40%			60%		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	α_1	SWAM	-0.08	-0.08008	0.0937	0.00173	-0.08008	0.0937	0.00221	-0.08003	0.0363	0.00276
Y_{2jkl}	β_1	SWAM	5	5.00366	0.0732	0.06502	5.00914	0.1828	0.0687	4.99833	0.0334	0.10171
Y_{3jkl}	λ_1	SWAM	5	5.00331	0.0662	0.06191	5.00607	0.1214	0.0714	4.99134	0.1732	0.09865
	λ_2	SWAM	-0.08	-0.08013	0.1587	0.00202	-0.07980	0.2550	0.0022	-0.08022	0.2687	0.00274
Y_{4jkl}	τ_1	SWAM	-0.008	-0.01030	**28.7500**	0.04583	-0.01040	**30.0000**	0.0028	-0.00935	**16.8750**	0.00332
	τ_3	SWAM	-0.025	-0.0319	**27.6117**	0.0603	-0.02218	**11.2910**	0.0801	-0.023	**7.9766**	0.10601

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; SWAM=Sampling Weight Adjusted for Missing Data

Table 5.12: Parameter estimates from weighted multilevel models using sampling selection weight adjusted for missing data (SWAM) from the Missing At Random missing mechanism with missing rates of 20% , 40% and 60% for data simulated from age -only, sex-only, sex-age, multivariable covariate models

SCALING METHOD A – MISSING AT RANDOM (MAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	α_1	SWAM	-0.08	-0.0799	0.1237	0.0020	-0.0800	0.0588	0.0023	-0.0799	0.1465	0.0030
Y_{2jkl}	β_1	SWAM	5	4.9952	0.0955	0.0678	4.9897	0.2056	0.0778	5.0205	0.4091	0.1053
Y_{3jkl}	λ_1	SWAM	5	5.0018	-0.0350	0.0863	5.0108	0.2168	0.0126	4.9960	0.0807	0.0004
	λ_2	SWAM	-0.08	-0.0804	0.5375	0.0019	-0.0800	0.0125	0.0020	-0.0799	0.0776	0.0030
Y_{4jkl}	τ_1	SWAM	-0.008	-0.0080	0.3450	0.0023	-0.0081	0.7158	0.0028	-0.0075	6.2853	0.0033
	τ_3	SWAM	-0.025	-0.0319	**27.6117**	0.0603	-0.0220	**11.8142**	0.0795	-0.0230	**8.6680**	0.1060

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; SWAFMD=Sampling Weight Adjusted for Missing Data

Table 5.13: Parameter estimates from weighted multilevel models using sampling weight adjusted for missing data (SWAM) from the Missing NOT At Random missing mechanism with missing rates of 20% , 40% and 60% for data simulated from age -only, sex-only, sex-age, multivariable covariate models

SCALING METHOD A – MISSING NOT AT RANDOM (MNAR)											
				20%			**40%**			**60%**	
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	α_1	SWAM	-0.08	-0.0801	0.1363	0.0801	-0.0794	0.7125	0.0022	-0.08003	0.0363	0.0801
Y_{2jkl}	β_1	SWAM	5	4.9982	0.0368	0.0644	5.0118	0.2352	0.0748	4.99833	0.0334	0.1017
Y_{3jkl}	λ_1	SWAM	5	4.9982	0.0368	0.0644	5.0118	0.2352	0.0748	4.99833	0.0334	0.1017
	λ_2	SWAM	-0.08	-0.0801	0.1363	0.0801	-0.0794	0.7125	0.0795	-0.08003	0.0363	0.0801
Y_{4jkl}	τ_1	SWAM	-0.008	-0.0080	**0.3450**	0.0458	-0.0081	**0.7158**	0.0028	-0.00935	**16.8750**	0.0033
	τ_3	SWAM	-0.025	-0.0319	**27.6117**	0.0679	-0.0220	**11.8142**	0.0824	-0.02301	**7.9766**	0.1085

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; SWAFMD=Sampling Weight Adjusted for Missing Data

Table 5.14: Parameter estimates from weighted multilevel models using sampling weight adjusted for missing data (SWAM) from the Not Missing At Random missing mechanism with missing rates of 20% , 40% and 60% for data simulated from age -only, sex-only, sex-age, multivariable covariate models

SCALING METHOD B – MISSING AT RANDOM (MAR)												
				20%			40%			60%		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	α_1	SWAM	-0.08	-0.0799	0.1738	0.0019	-0.0800	0.0025	0.0022	-0.0799	0.1375	0.0029
Y_{2jkl}	β_1	SWAM	5	4.9975	0.0508	0.0666	4.9890	0.2196	0.0713	5.0215	0.4300	0.1034
Y_{3jkl}	λ_1	SWAM	5	5.0041	0.0816	0.0858	5.0077	0.1536	0.0133	4.9973	0.0550	0.0006
	λ_2	SWAM	-0.08	-0.0805	0.5675	0.0019	-0.0800	0.0063	0.0021	-0.0799	0.1150	0.0029
Y_{4jkl}	τ_1	SWAM	-0.008	-0.0079	1.7050	0.0021	-0.0080	0.3049	0.0025	-0.0078	1.9236	0.0033
	τ_3	SWAM	-0.025	-0.0183	26.6028	0.0696	-0.0288	15.1637	0.0802	-0.0217	13.0408	0.0929

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; SWAFMD=Sampling Weight Adjusted for Missing Data

Table 5.15: Parameter estimates from weighted multilevel models using sampling weight adjusted for missing data (SWAM) from the Not Missing At Random missing mechanism with missing rates of 20% , 40% and 60% for data simulated from age -only, sex-only, sex-age, multivariable covariate models

SCALING METHOD B -MISSING NOT AT RANDOM (MNAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	α_1	SWAM	-0.08	-0.0801	0.1075	0.0801	-0.0797	0.3725	0.0797	-0.0799	0.1137	0.0800
Y_{2jkl}	β_1	SWAM	5	5.0040	0.0794	0.0658	5.0105	0.2090	0.0694	5.0016	0.0328	0.1047
Y_{3jkl}	λ_1	SWAM	5	5.0032	0.0638	0.0610	5.0028	0.0552	0.0746	4.9940	0.1192	0.0974
	λ_2	SWAM	-0.08	-0.0800	0.0613	0.0019	-0.0797	0.3475	0.0022	-0.0801	0.1662	0.0027
Y_{4jkl}	τ_1	SWAM	-0.008	-0.0079	0.8087	0.0434	-0.0080	0.4529	0.0027	-0.0094	**16.8750**	0.0033
	τ_3	SWAM	-0.025	-0.0309	23.7023	0.0613	-0.0195	22.1277	0.0796	-0.0230	**7.9766**	0.1060

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; SWAFMD=Sampling Weight Adjusted for Missing Data

Chapter 6

Comparative Analysis of Survey Data with emphasis on Item non-response Weighting Method

Chapter Outline

The aim of this thesis is to assess weighting methods from different scenarios in multilevel modelling of a continuous outcome with missing data. Sequel to Chapter 6, five weighting adjustment methods were assessed. In Chapter 6, the focus is on the application of item non-response weight to the multilevel modelling of annual household expenditure of food purchased away from home in Jamaica and also the modelling of individual reported income in China.

6.1 Aim and Scope of the Analysis

Statistical analysis and methodology involved in the analysis of complex and large survey data are a challenge to data analysts, especially as they relate to datasets from non-uniform sampling technique. The creation of models and the development of interpretations from hierarchical survey datasets continue to generate debate (Maas & Hox, 2004). Accordingly, the aim of Chapter 6 was to develop a series of weighted multilevel models for continuous outcome variables with missing data from multistage survey datasets, using the item non-response weight in the analysis. In these analyses, the lessons learnt in Chapter 5 from the simulation study regarding the construction and incorporation of item non-response weights are used. Specifically, the item non-response weight will be applied at different stages of the model to illustrate its effects on the model-selection processes. The Jamaica Survey of Living Conditions 2007 (JSLC 2007) data is used in the first analysis while the China Health and Nutrition Survey (CHNS) data for 1989 and 2011 is used in the second analysis. Using the JSLC 2007, the analysis focused on the reported household annual expenditure of meals purchased away from home while the analysis for the CHNS datasets focused on the reported individual income.

The set of analyses in this chapter stem from the simulation study conducted in Chapter 5 on the reliability of weighting methodologies for survey data with missing data from a multistage survey design. The findings from the simulation study in Chapter 5 confirmed that the item non-response weight was the most reliable under different proportions of missing data, hence the application to the modelling of food-related expenditure and income variables.

.

6.2 Objectives and Rationale for the Analysis

The first objective of the analysis was the creation of two- and three- level random coefficient weighted models. The focus of these models was the annual household expenditure data for meals purchased away from home in the JSLC 2007. The second objective of the analysis involved the creation of a four-level weighted model at the province, stratum, community, and household levels for the reported individual income data in the CHNS for the years 1989 and 2011.

The rationale for these analyses were three-fold. The first rationale stemmed from the findings of the simulation study in Chapter 5, while the second rationale was to investigate the size of the intraclass correlation and the need for multilevel model. Huang (2018) affirmed that it is a myth to state that a multilevel model would be necessary on the basis of the intraclass correlation. This misconception will be investigated by varying the analysis using different models to highlight different intraclass values and estimated parameters. The third rationale for the analyses in this Chapter is to compare weighted parameter estimates from different multilevel models when estimates are adjusted for item non-response. Furthermore, the gains of using item non-response weight as an adjustment method to compensate for missing data when developing a multilevel model will be elaborated upon.

6.3 Multilevel Models in the Analysis with 14 Parishes

The proposed study models were multilevel as they were developed to fully incorporate the hierarchical structures in the survey design, and also to achieve the chapter's objectives. The method of parameter estimation is described in Chapter 4.

6.3.1 JSLC2007 Multilevel Models

Model II-A

Two multilevel model scenarios were investigated in this analysis. The first scenario involved a series of two-level models consisting of Parish and Household levels with the application of the item non-response weight at the Parish level. I started the analysis with the random effect as expressed in 6.1, then followed by the fixed effect version with the addition of the Parish as predictor variable in 6.1.1. These models are labelled as Model II-A in the illustrations in Figure 1 as well as in the result in Table 6.3.

$$y_{ji} = x_{ji}^T \beta + u_j + \varepsilon_{ji} \qquad (6.1)$$

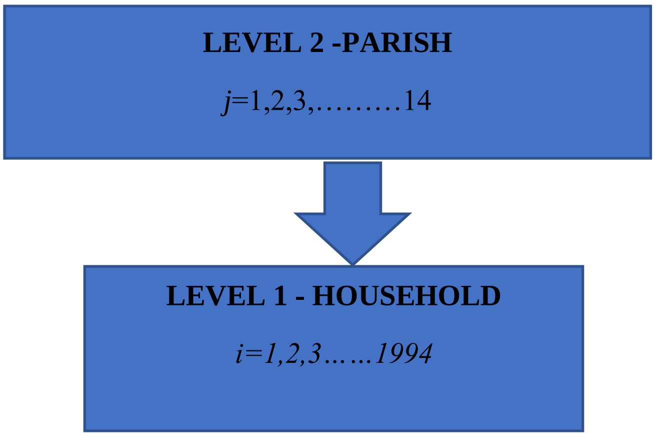
$$y_{ji} = x_{ji}^T \beta + d_j + \varepsilon_{ji} \qquad (6.1.1)$$

In both models y_{jl} is the dependent variable, the log of the annual household expenditure of meals purchased away from home for the i^{th} household in the j^{th} Parish. In model 6.1, the random and error terms are represented by $u_j \sim \mathrm{N}(0, \sigma_j^2)$, $\varepsilon_{ji} \sim \mathrm{N}(0, \sigma_{ji}^2)$ respectively while in the 6.1.1 represented fixed term for the Parish was represented by d_j. The x_{ji}^T denoted the vector of subject level variables from the socio-demographic (Age, Household Size, Sex, Education, Employment Status, and Occupation Status). The result of the two-level model series was denoted as model II-A, and are index from A1 to A8 in Table 6.3 followed by the weighted fixed effect model with the inclusion of the parish as one of the explanatory variables. The result of this models respectively and illustrated in Table 6.3.1. denoted Mode II-A9.1 and

Model II-A9.2 respectively to highlight the building process and final models. Further collapsing of the parish into two categories resulted in results in Table 6.3.1.1.

In the parameter estimation, w_{ji}^{1} is incorporated as the item non-response weight from the response probability of the household and average per Parish representing non-response for each individual in the Parish. The incorporation is done by weighting the parameter estimator by w_{ji}^{1} to adjust for non-response. Figure 6.1 illustrates the two-level model denoted Model II-A.

Figure 6.1. Model II-A



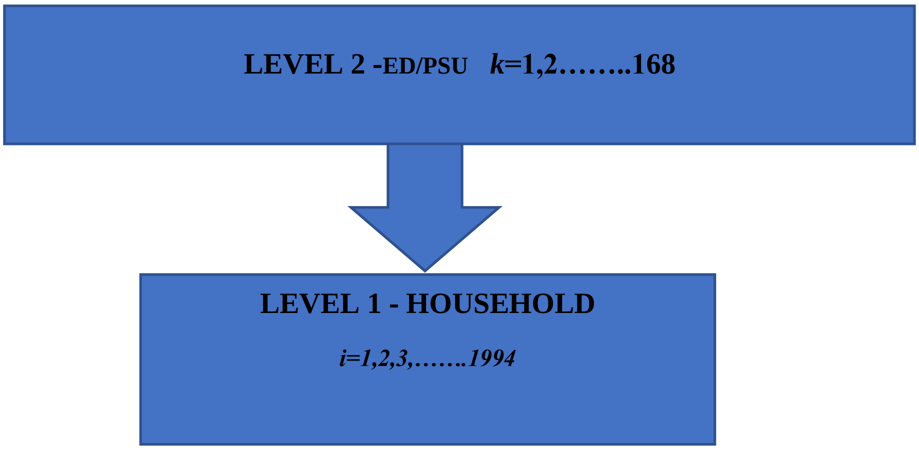
Model II-B

In continuation of the two-level model, an alternative two-level model consisting of the Enumeration District (ED), the Primary Sampling Unit (PSU) and Household was also developed for comparison purposes. There are 168 PSUs which was too many to consider for a fixed effect only model. The random effect model is stated in 6.2.

$$y_{ki} = x_{ki}^{T}\beta + v_k + \varepsilon_{ki} \qquad (6.2)$$

In this model y_{ki} represent the log of the annual household expenditure of meals purchased away from home for the i^{th} household in the k^{th} PSU. In this model, $v_k \sim \mathrm{N}(0,\sigma_k^2)$ and $\varepsilon_{ki} \sim \mathrm{N}(0,\sigma_{ki}^2)$ represent the random components at the ED and household levels, respectively. In this model, w_{ki}^2 represented the item non-response weight from the response probability of the household and average per ED unlike the earlier model in 6.1 which is average per Parish. This weight is used to adjust the estimates for item non-response for each individual at the ED level. In this model II-B and illustrated in Figure 6. 2.

Figure 6. 2. Model II-B



LEVEL 2 -ED/PSU *k*=1,2……..168

LEVEL 1 - HOUSEHOLD

i=1,2,3,…….1994

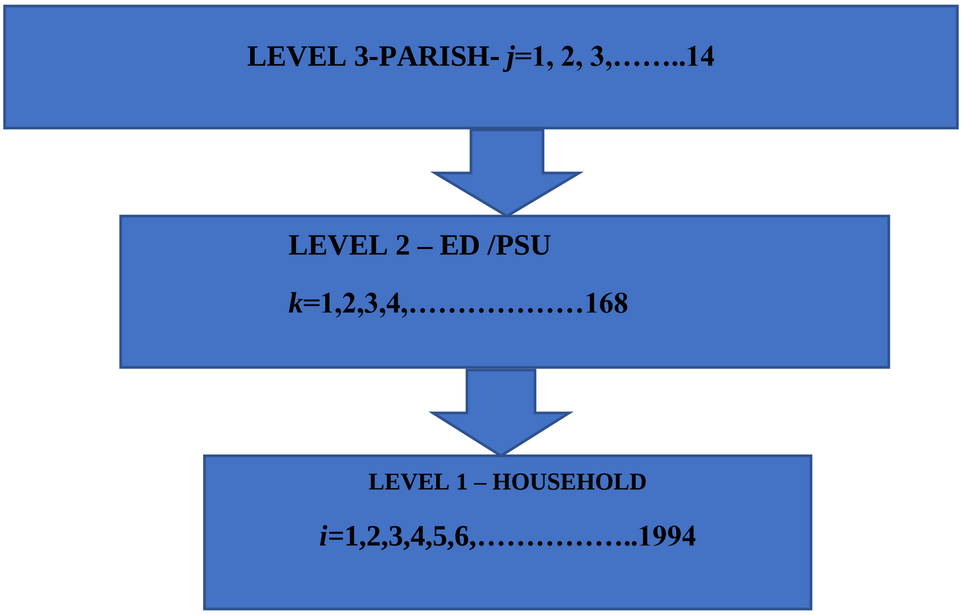
Model -III A

The second scenario involves the development of the final model in the JSLC 2007 analysis, a three-level model to draw comparisons with the two-level models in the first scenario. This model incorporated the Parish, ED, and Household levels while using the item non-response weight previously averaged at the ED level. The three-level model represented as:

$$y_{jki} = x_{jki}^T \beta + u_j + v_{jk} + \varepsilon_{jki} \qquad (5.3)$$

where $u_j \sim \mathrm{N}(0,\sigma_j^2)$, $v_{jk} \sim \mathrm{N}(0,\sigma_{jk}^2)$ and $\varepsilon_{jki} \sim \mathrm{N}(0,\sigma_{jki}^2)$ represents the random component at the Parish, ED, and household level respectively. The model is illustrated in Figure3.

Figure 6. 3. Model III-A


LEVEL 3-PARISH- *j*=1, 2, 3,……..14
LEVEL 2 – ED /PSU
k=1,2,3,4,………………168
LEVEL 1 – HOUSEHOLD
i=1,2,3,4,5,6,……………..1994

6.3.2 Multilevel Model for the China Health and Nutrition Survey (CHNS)

For the CHNS analyses, two separate multilevel models are formulated for comparative analysis. A four – level model was developed for each Year 1989 and 2011 respectively to illustrate the hierarchical structure of the data in the unweighted and weighted models. Consequently, the process was repeated with a three – level models each year for comparative analysis to determine if the use of three level will cause a change in the predictor variables and parameter estimates.

Four Level Model IV

The four- level model is grounded on the basis that the household is nested within the Neighborhood-Village which is nested within county-city and nested within the province.

The four-level model for the CHNS 1989 and 2011 is illustrated as follows:

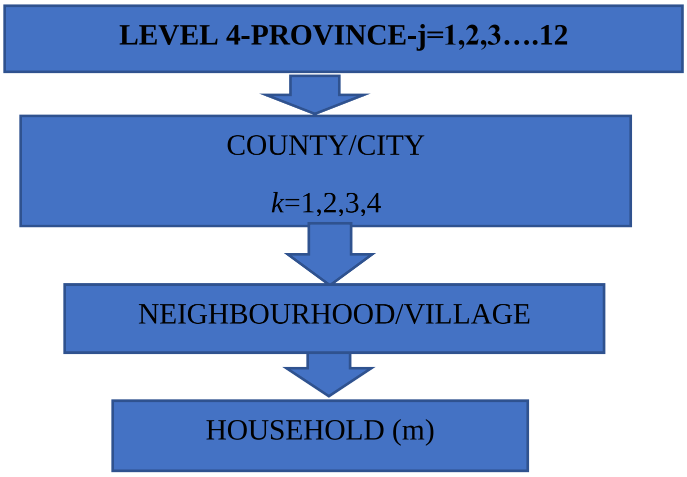
Let j represent the provinces in each model, the highest level. In 1989, data was collected from 8 Provinces compared with the 12 in the 2011 data set. The provinces are the highest level in the model. Hence for every j , there are k nested County/City further nested by l Neighbourhood -Village and m individuals at the lowest level. (Zhang et al, ,2014).

In the four-level model IV, the outcome variable y_{ji} was defined as the reported total income from i^{th} household nested within l^{th} Neighbourhood -Village then nested within k^{th} County-City then nested within the j^{th} Province in model **IV**

$$y_{jkli} = x^{T}_{jklm}\beta + u_j + v_{jk} + \tau_{jkl} + \varepsilon_{jkli} \qquad (6.4)$$

The x^T is the vector variables for age, gender, marital status, and completed years of formal education. The notation: $u_j \sim \mathrm{N}(0,\sigma^2_j)$, $v_{jk} \sim \mathrm{N}(0,\sigma^2_{jk})$ $\tau_{jkl} \sim \mathrm{N}(0,\sigma^2_{jkl})$ and $\varepsilon_{jkli} \sim \mathrm{N}(0,\sigma^2_{jkli})$ represents the random components at the Province, County – City, Neighbourhood-Village, and household, respectively as illustrated in Figure 6. 4 **Model IV**

Figure 6.4. Model IV-A



Three -Level Model III for the Year 1989 and 2011

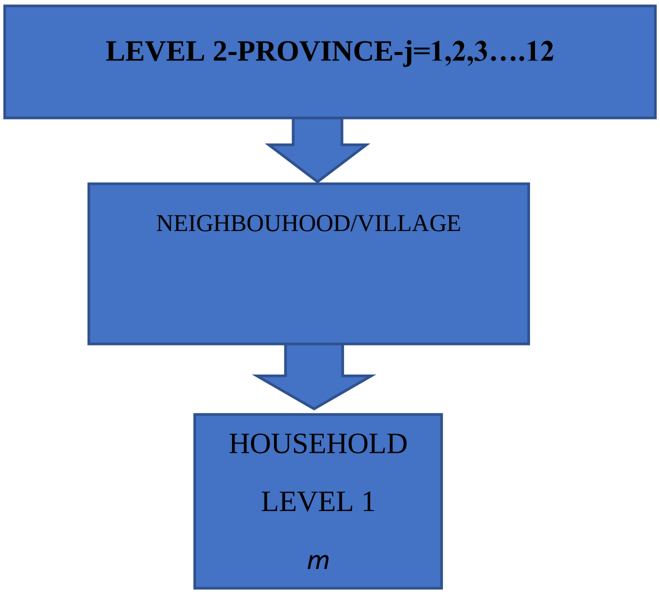
In the three -level model, the Neighborhood -Village is nested within the province. The three -level model for the CHNS 1989 & 2011 is illustrated as follows:

Let j represent the provinces for $j = 1, 2, 3.................12.$, for the twelve Provinces for the Year 2011 survey. Within these Provinces, l, Neighbourhood-Village is selected for $l = 1, 2, 3....161.$ consisting of m individuals with $m = 1, 2, 3,.....2109.$ In the three-level model, y_{jlm}, the outcome variable reported income from m^{th} household nested within l^{th} Neighbourhood-Village nested within the j^{th} Province is presented as follows:

$$y_{jli} = x_{jli}^{T}\beta + u_{j} + v_{jl} + \varepsilon_{jli} \qquad (6.5)$$

The x^{T} is the vector variables for age, gender, marital status, and completed years of formal education. The notation: $u_{j} \sim \mathrm{N}(0, \sigma_{j}^{2})$, $v_{jl} \sim \mathrm{N}(0, \sigma_{jl}^{2})$ and $\varepsilon_{jli} \sim \mathrm{N}(0, \sigma_{jli}^{2})$ represents the random components at the Province, Neighbourhood-Village, and household, respectively as illustrated in Figure 6.5: **Model III**

Figure 6.5: Model III



6.4 Estimating Response Probabilities for each Data Set and Item Non-Response Weight

In the development of the item non-response weight, the response probability via logistic regression was used to determine the probability of every individual's likelihood of responding to the outcome variable. Dehija and Wahba (1999) and Kalton and Flores-Cervantes (2003) applied logistic regression in the estimation of response probability and also used the inverse of the probability to create item non-response weight. In these analyses, the notation and context of the logistic regression is described in the section below:

For each dataset let R_i be the missing indicator variables for every i individual, and also R_i is a binary variable containing: 1 for missing, 0 for not missing. Similarly, socio-demographic variables of the individuals are represented by X_i, a vector of all the independent variables under the assumptions of full data.

The regression model used in the study and in Bethlehem et al. (2011):

$$\log\left(\frac{\mathrm{R_i}}{1\text{-}\mathrm{R_i}}\right)=\mathrm{logit(p(X_i))=X_i'\beta} \qquad (6.6)$$

the final model is determined for only significant (p-value less than 0.05) variables are included in the model. The response probability, $\hat{p}_i(R)$ is finally estimated as follows:

$$\hat{p}_i(R_i)=\frac{\exp(\mathrm{X}_i'\,\beta)}{1+\exp(X_i'\beta)} \qquad (6.7)$$

In the model, the inverse of $\hat{p}_i(R_i)$ is used to determine the weight for all the models.

6.5 Intraclass Correlation Estimation

The Intraclass Correlation (ICC) was denoted as ρ and defined as the measure of the proportion of variation in the outcome variable that occurred between groups in relation to the total variation present (Finch et al., 2014). In these analyses, the proportion of variance at each level of the data hierarchy were estimated to investigate underlying relationship between the selected model and the intraclass estimated correlations. Below is an illustration of the estimation of ICC for a three-level model with variances at household (σ_ε^2), ED (σ_v^2) and Parish (σ_u^2)levels:

ICC for the household level variation:

$$\rho_\varepsilon = \frac{\sigma_\varepsilon^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2} \qquad (6.8)$$

ICC for the PSU level variation:

$$\rho_v = \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2} \qquad (6.9)$$

ICC for the Parish level variation:

$$\rho_u = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2} \qquad (6.10)$$

6.6 Model Building and Selection Approach

The models in these analyses were created using a forward selection approach while noting the level of significance as the variables were added to the model. The baseline for adding a variable was a p-value less than 0.05. Any variable with a p-value > 0.05 was excluded from the model. Additionally, the effect of the weights with regards to the Akaike Information Criterion (AIC) from the pseudo-maximum likelihoods were also observed. Hirotugu Akaike(1978) introduced an information criterion which is defined as :

AIC = (-2) log(maximum likelihood) + 2(number of parameters)

6.7 Analysis of the JSLC 2007

The JSLC 2007 was conducted from May 1 to July 31, 2007 as part of a longitudinal study on the social conditions of the Jamaican population as explained in Chapter 3. The survey design employed multistage sampling for the sample selection. The island of Jamaica is divided into 14 Parishes with 5, 235 Enumeration Districts (EDs) proportional to the size of the population in each Parish. In the JSLC 2007, 168 EDs were sampled from the available 5,235 EDs. The size of the sampled EDs per parish ranged from 6 for the smaller to 36 for the larger parishes. Table 5.1 contains the sampled EDs per parish, for example, in the parish of Kingston, 6 EDs or Primary Sampling Units were sampled out of the 224 Enumeration Districts. In the survey design, 16 households were sampled in each of the EDs with the expectation that the total responses would be 2,688 households from the sampled units. However, at the end of the survey, only 1,994 households responded, of which 1,653 households responded to the outcome variable. The minimum number of households in a given ED was 3 while the maximum was 16 thus affirming the contrast between the expected and actual sample size. The Unit non-response adjustment is not considered in this study because there is no availability of information on the sampled households who did not respond during the survey and adjusted such as the post -stratification method is not the focus of this study. In addition, the focus of this Chapter is the application of the item non-response weight which was identified as the least bias in the simulation study.

Table 6.1: Distribution of the sampled EDs per Parish

PARISH NUMBER	PARISH	SAMPLE PSU	TOTAL PSU
1	KINGSTON	6	224
2	ST. ANDREW	36	973
3	ST. THOMAS	6	224
4	PORTLAND	6	187
5	ST. MARY	6	278
6	ST. ANN	10	319
7	TRELAWNY	6	178
8	ST. JAMES	10	347
9	HANOVER	6	159
10	WESTMORELAND	8	322
11	ST. ELIZABETH	10	308
12	MANCHESTER	12	343
13	CLARENDON	14	467
14	ST. CATHERINE	32	906
TOTAL		**168**	**5,235**

The motivation for the selected variable in the JSLC 2007 analysis was based on the study from Manrique and Jensen (1998). Prior exploratory analysis of the relationship between the socio-demographics and the outcome variable. The outcome variable in the JSLC 2007 had 17.35% missing of the 1994. Manrique and Jensen (1998) reported that food away from home was a function of time, wage, and vector of socio-demographic variables such as age, gender, race, and household size. Similar variables were postulated by Bai et al. (2010), Lee (2006), McCracken and Brandt (1987), and Ma et al. (2006). The covariates in this analysis were extracted from the socio-demographic variables for each head of household in the JSLC 2007.

The variables selected included age, household size, sex, level of education, employment, and occupational statuses, respectively. The average age of the household heads was 48.97 years with standard deviation of 16.73 years. The youngest head of household was aged 16 while the oldest was aged 99. Household size ranged from 1 to 18 members with an average of 3.4 members.

In the JSCL 2007, there were more male-headed households (53.66%; n=1,070) than female-headed households (46.34%; n=924). For the analysis, the education variable was collapsed into two categories: educated (8.81%; n=176) and neither stated nor had a formal education (91.17%; n=1,818). Three occupation categories were used in the model development: Office Related Work; Non-Office Related Work and Non-Classified (ORW= 14.74%, n=294; NORW=69.45%, n=1385; NC=15.80%, n=315) and two employment categories(unemployed and not classified =15.59%; n=311; employed= 84.40%; n=1,683).

Prior to the weighted model, the item non-response was derived as the inverse of the response probability as a function of age of the respondents. This was based on the result of the logistic regression model for the JSLC 2007. The analysis revealed that of all the socio-demographic variables, age was found to be a significant ($p<0.05$) predictor of the missing indicator variable.

6.7.1 Result of the JSLC2007 Multilevel Model Analysis

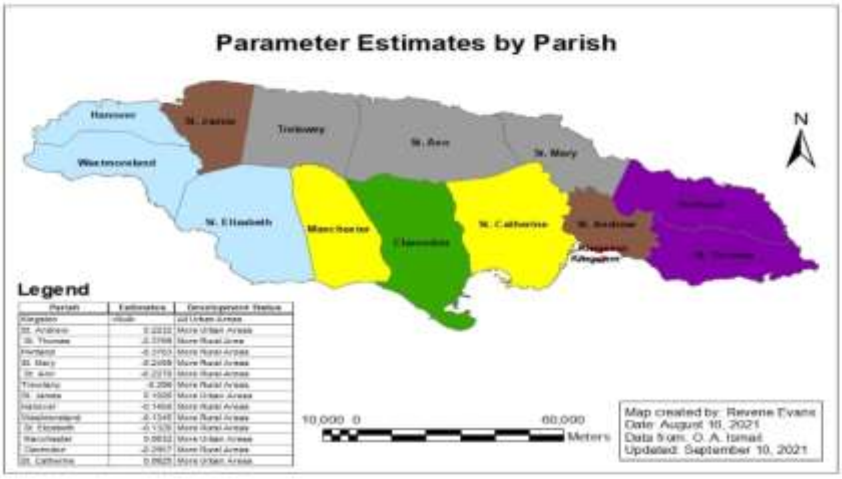
In the JSLC 2007 analysis, two separate weighted two-level models were investigated. Model II-A seeks to illustrate the effects of the item non-response weight on the parameter estimate when a two-level model only consists of the Parish and the household level random slope. In this model, the minimum item non-response weight for a Parish was 1.184 while the maximum was 1.252. Similarly, the sample size per parish ranges from 65 to 404. The item non-response weight in this scenario represents the weighted estimate at the Parish level which manifested in the estimated parameters in Models II-A1 to Model II-A8 in Table 6.3. A continual decrease of the AIC values was observed as more variables are added to the model, until the final and best Model II-A6 consisting of age, household size level of education and occupation as the significant predictor variables ($p< 0.05$).

In order to make comparison with the first-set of analysis involving the Parish as the second and highest level in ModelII-A1 to ModelII-A8. Another two -level ModelII-B1 to ModelII-B8 having the ED as the highest level instead of Parish were also developed. There are 168 Enumeration Districts in the model compared to the earlier model with 14 Parishes. In these Models, the item non-response weight constant for the ED as a weight to adjust for missing data from household heads in each of the selected EDs. In this scenario, the average item weight per PSU ranges from 1.109 to 1.357 while the sample size per ED also ranges from 5 to 18. These results show that for small sample size, the average weight per ED were higher than the average weight per Parish with larger sample. This phenomenon account for the estimates in Tables 6.3 and 6.4 respectively.

In ModelII-A9.1, ModelII-A9.2, ModelII-A9.1.1 & ModelII-A9.2.2 the Parish was added as a fixed effect, similar results were observed with a lower AIC value as demonstrated in Table 6.3.1. This suggested that the addition of the Parish did not affect parameter estimates significantly. However, the results revealed that the parameter estimates for the parishes in some less developed parishes especially those with higher number of rural areas have reduced annual expenditure of meals purchased away from home. Some of the parishes with lower annual expenditure on meals away from home are contiguous. The parishes with the predominantly rural areas include St. Thomas, Portland, St. Mary, St. Ann, Trelawny, Hanover, Westmoreland, St.

Elizabeth and Clarendon, while the parishes with more urban areas includes St. Andrew, St. James, Manchester and St. Catherine. These parishes have a greater annual household expenditure of meals purchased away from home. These parishes have some common qualities especially Kingston and St. Andrew which is the commercial hub of the island with the highest concentration of the tertiary institutions. The parish of St. Catherine is in close proximity to these two Parishes. St. James is the tourism hub of the island which also contributed to the development of the parish. Similar qualities exist in the parish of Manchester which has a university and opportunity for employment in small and medium scale industries as illustrated in Figure 6.

Figure 6.: Illustration of the parameter estimates from the fixed effect model



The parishes in grey while the baseline is Kingston in red below St. Andrew in Figure 6 have higher concentration of fast-food restaurants when compared to the more rural dominated parishes in the eastern and western end.. This clearly suggest that the tendency of the household members to eat away from home will be higher for persons who reside in the parishes with higher urban areas than those who reside in more rural areas of the island. Further merging of these parishes with negative parameter estimates as a group and those with positive parameter estimates clearly revealed the differences in the parameter estimates in the fixed effects in Table 6.4.

Unlike the estimates in Table 6.3, the parameter estimates for the ED scenario were found to be higher as demonstrated in Table 6.6. Similar variables such as age, household size, level of education and occupation were the significant predictor variable in the best model II-B8. In the three-level model involving EDs nesting in the Parish. The lessons learnt from these analyses, is that sample size determines the magnitude of the weight per unit (Parish or ED) because the sizes of each cluster (Parish or ED). Also, the magnitude of the parameter estimates is proportional to the size of the weight. Additionally, level at which the weight is applied is more relevant in the parameter estimate than the hierarchical levels in the model.

Table 6.2 also revealed that the majority of the proportion of the variation in the outcome variable were found in the household level of the models. However, as the model level increased from two to three, the ICC values increased for the Parish and ED while decreasing for the household level.

Table 6.2: Estimated ICC Values for the Levels in the Models in the JSLC 2007

Model	**Variance estimates and ICC at each level**
Model II-A $y_{jl} = x_{jl}^T\beta + u_j + \varepsilon_{jl}$ *, for* $\sigma_{u_j}^2$ *and* $\sigma_{\varepsilon_{jl}}^2$ *represents Parish and Household variances respectively*	σ_u^2 = 0.039 σ_ε^2 = 0.868 ρ_{PARISH} = 0.043 $\rho_{HOUSEHOLD}$ = 0.957
Model II-B $y_{kl} = x_{kl}^T\beta + v_k + \varepsilon_{kl}$ *,* *for* $\sigma_{V_k}^2$ *and* $\sigma_{\varepsilon_{kl}}^2$ *represents ED and household variances respectively*	σ_v^2 = 0.121 σ_ε^2 = 0.795 ρ_{ED} = 0.132 $\rho_{HOUSEHOLD}$ = 0.868

Model III-A $y_{jkl} = x_{jkl}^T \beta + u_j + v_{jk} + \varepsilon_{jkl}$ for $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$ and $\sigma^2_{\varepsilon_{jkl}}$ *represents Parish, ED and Household variances respectively.*	σ^2_u = 0.035 σ^2_v = 0.080 σ^2_ε = 0.795 ρ_{PARISH} = 0.039 ρ_{ED} = 0.088 $\rho_{HOUSEHOLD}$ = 0.873

Table 6.3: Parameter Estimates for Two-Level Random Coefficient Model using Item Non-response Weight Average at the Parish level

Variable	Model II-A1		Model II-A2		Model II-A3		Model II-A4		Model II-A5		Model II-A6		Model II-A7		Model II-A8	
	PE (SE)	p-value	PE (SE)	p-value	PE (SE)	p-value	PE (SE)	p-value	PE (SE)	p-value	PE (SE)	p-value	PE (SE)	p-value	PE (SE)	p-value
Fixed Effects																
Constant	10.8948 (0.0609)	0.0000	11.2296 (0.07916)	0.0000	10.8859 (0.1066)	0.0000	10.8930 (0.1102)	0.0000	10.8289 (0.1066)	0.0000	10.4354 (0.1318)	0.0000	10.4542 (0.1256)	0.0000	10.4289 (0.1162)	0.0000
Age			-0.0070 (0.0013)	0.0000	-0.0091 (0.0012)	0.0000	-0.0092 (0.0012)	0.0000	-0.0084 (0.0012)	0.0000	-0.0071 (0.0014)	0.0000	-0.0079 (0.0014)	0.0000	-0.0078 (0.0149)	0.0000
Household Size					0.1221 (0.0134)	0.0000	0.1222 (0.0133)	0.0000	0.1232 (0.1344)	0.0000	0.1156 (0.0144)	0.0000	0.1155 (0.0148)	0.0000	0.1154 (0.0149)	0.0000
Sex																
Female							-0.0183 (0.0347)	0.5990	-0.0172 (0.0343)	0.6160	-0.0249 (0.0346)	0.4720	-0.0187 (0.0339)	0.5820		
Male -Ref																
Education																
Educated									0.3732 (0.0473)	0.0000	0.3215 (0.0426)	0.0000	0.2486 (0.0353)	0.0000	0.2475 (0.0361)	0.0000
Not Stated – Ref																
Employment Status																
Employed											0.4169 (0.0869)	0.0000	-0.2020 (0.1137)	0.0760		
Unemployed-Ref																
Occupation																
Office Related Work (ORW)													0.84020 (0.0829)	0.0000	0.6512 (0.0787)	0.0000
Non-Office Related Work (NORW)													0.6101 (0.1203)	0.0000	0.4195 (0.0929)	0.0000
Not -Classified -Ref																
Random Component																
σ^2_{PARISH}	0.038730		0.035872		0.034857		0.034782		0.029207		0.02955		0.02739		0.027889	
$\sigma^2_{HOUSEHOLD}$	0.868997		0.857106		0.786592		0.786592		0.775809		0.758467		0.747706		0.748052	
AIC	5416.97		5390.78		5222.07		5223.86		5196.44		5153.99		5128.59		5125.97	

PE=Parameter Estimates, SE = Standard Error, AIC = Akaike Information Criterion

Table 6.4: Parameter Estimates for Two-Level Random Coefficient Model using Item Non-response Weight Average at the Parish level incorporating Parish as a fixed effect variable

	Model II-A9.1			**Model II-A9.2**		
Variable	**Parameter Estimate**	**(S.E)**	**p-value**	**Parameter Estimate**	**(S.E)**	**p-value**
Constant	10.5646	(0.1082)	0.0000	10.5469	(0.0988)	0.0000
Age	-0.0078	(0.0014)	0.0000	-0.0077	(0.0014)	0.0000
Household Size	0.1154	(0.0148)	0.0000	0.1153	(0.0149)	0.0000
Sex						
Female	-0.0169	(0.0339)	0.6180			
Male-Ref						
Education						
Educated	0.2342	(0.0367)	0.0000	0.2333	(0.0374)	0.0000
Not Stated-Ref						
Employment Status						
Employed	-0.1832	(0.1175)	0.1190			
Unemployed-Ref						
Occupation						
ORW	0.8196	(0.0848)	0.0000	0.6484	(0.0794)	0.0000
NORW	0.5919	(0.1232)	0.0000	0.4191	(0.0948)	0.0000
Not-Classified-Ref						
Parish						
St. Andrew -2	0.2232	(0.0087)	0.0000	0.2201	(0.0082)	0.0000
St. Thomas -3	-0.3789	(0.0134)	0.0000	-0.3879	(0.0132)	0.0000
Portland – 4	-0.3703	(0.0133)	0.0000	-0.3758	(0.0113)	0.0000
St. Mary -5	-0.2459	(0.0116)	0.0000	-0.2493	(0.0106)	0.0000
St. Ann – 6	-0.2278	(0.0201)	0.0000	-0.2389	(0.0152)	0.0000
Trewlany -7	-0.2990	(0.0149)	0.0000	-0.3068	(0.0118)	0.0000
St. James -8	0.1026	(0.0116)	0.0000	0.0978	(0.0094)	0.0000
Hanover – 9	-0.1458	(0.0153)	0.0000	-0.1500	(0.0137)	0.0000
Westmoreland -10	-0.1345	(0.0120)	0.0000	-0.1371	(0.0110)	0.0000
St. Elizabeth -11	-0.1326	0.0137	0.0000	-0.1366	(0.0122)	0.0000
Manchester – 12	0.0632	0.0174	0.0000	0.0592	(0.0160)	0.0000
Clarendon -13	-0.2957	0.0142	0.0000	-0.3015	(0.0116)	0.0000
St. Catherine -14	0.0625	0.0152	0.0000	0.0572	(0.0131)	0.0000
Parish -1 -Ref	-	-	-	-	-	-
σ^2_{PARISH}	0.0000			0.0000		
$\sigma^2_{HOUSEHOLD}$	0.7421			0.7417		
AIC	**5084.1**			**5081.2**		

Table 6.5: Parameter Estimates for Two-Level Random Coefficient Model using Item Non-response Weight Average at the Parish level incorporating Parish as a fixed effect variable

	Model II-A9.1.1			Model II-A9.2.2		
Variable	Parameter Estimate	(S.E)	p-value	Parameter Estimate	(S.E)	p-value
Constant	10.5744	(0.1124)	0.0000	10.5543	0.1032	0.0000
Age	-0.0080	(0.0015)	0.0000	-0.0078	0.0015	0.0000
Household Size	0.1145	(0.0336)	0.0000	0.1144	0.0155	0.0000
Sex						
Female	-0.0181	(0.0336)	0.5880			
Male-Ref						
Education						
Educated	0.2453	0.0282	0.0000	0.2443	0.0289	
Not Stated-Ref						
Employment Status						
Employed	-0.2143	0.1191	0.0720			
Unemployed-Ref						
Occupation						
ORW	0.8479	(0.0849)	0.0000	0.6472	(0.0796)	0.0000
NORW	0.6222	(0.1260)	0.0000	0.4199	(0.0964)	0.0000
Not-Classified-Ref						
Parish						
d281214	0.1224	(0.0464)		0.1172	(0.0462)	0.0110
d567910111334	-0.2419	(0.0344)		-0.2489	(0.0324)	0.0000
Parish -1 -Ref	-	-	-	-	-	-
σ^2_{PARISH}	0.0011			0.0012		
$\sigma^2_{HOUSEHOLD}$	0.7466			0.7470		
AIC	5105.94			5103.45		

Table 6.6: Parameter Estimates for Two-Level Random Coefficient Model using Item Nonresponse Weight Average at the ED Level

	Model II-B1		Model II-B2		Model II-B3		Model II-B4		Model II-B5		Model II-B6		Model II-B7		Model II-B8	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects																
Constant	10.9860 (0.0352)	0.0000	11.3542 (0.0756)	0.0000	11.0039 (0.0786)	0.0000	11.0119 (0.0797)	0.0000	10.9307 (0.0816)	0.0000	10.5561 (0.1118)	0.0000	10.5669 (0.1108)	0.0000	10.5378 (0.1062)	0.0000
Age			-0.0078 (0.0014)	0.0000	-0.0099 (0.0013)	0.0000	-0.0099 (0.0013)	0.0000	-0.0091 (0.0013)	0.0000	-0.0078 (0.0014)	0.0000	-0.0085 (0.0014)	0.0000	-0.0084 (0.0014)	0.0000
Household Size					0.1246 (0.0105)	0.0000	0.1247 (0.0106)	0.0000	0.1249 (0.0106)	0.0000	0.1176 (0.0107)	0.0000	0.1169 (0.0108)	0.0000	0.1169 (0.0108)	0.0000
Sex																
Female							-0.0201 (0.0435)	0.6440	-0.0202 (0.0436)	0.6430	-0.0291 (0.0436)	0.5050	-0.0259 (0.0434)	0.5510		
Male -Ref																
Education																
Educated									0.3585 (0.0598)	0.0000	0.3163 (0.0809)	0.0000	0.2511 (0.0621)	0.0000	0.2507 (0.0624)	0.0000
Not Stated – Ref																
Employment Status																
Employed											0.3976 (0.0809)	0.0000	-0.2283 (0.1545)	0.1390		
Unemployed-Ref																
Occupation																
Office Related Work (ORW)													0.8394 (0.1617)	0.0000	0.6245 (0.0958)	0.0000
Non-Office Related Work (NORW)													0.6206 (0.1571)	0.0000	0.4043 (0.0788)	0.0000
Not -Classified -Ref																
Random Component																
σ^2_{PSU}	0.121243		0.117924		0.126309		0.126238		0.110822		0.109561		0.104006		0.104523	
$\sigma^2_{HOUSEHOLD}$	0.796199		0.782340		0.706104		0.705936		0.702076		0.686744		0.678646		0.679141	
AIC	5392.82		5358.74		5176.98		5178.72		5154.05		5113.38		5088.93		5086.90	

PE=Parameter Estimates, SE = Standard Error, AIC = Akaike Information Criterion

Table 6.7: Parameter Estimates for Three-Level Random Coefficient Model at Using Item Nonresponse Weight Average at the ED Level

	Model III-C1		Model III-C2		Model III-C 3		Model III-C4		Model III-C5		Model III-C6		Model III-C7		Model III-C8	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects																
Constant	10.8969 (0.0636)	0.0000	11.2468 (0.0802)	0.0000	10.8938 (0.1023)	0.0000	10.9013 (0.1049)	0.0000	10.8437 (0.1013)	0.0000	10.4612 (0.1285)	0.0000	10.4767 (0.1230)	0.0000	10.4500 (0.1139)	0.0000
Age			-0.0073 (0.0013)	0.0000	-0.0095 (0.0011)	0.0000	-0.0095 (0.0011)	0.0000	-0.0087 (0.0011)	0.0000	-0.0074 (0.0013)	0.0000	-0.0081 (0.0014)	0.0000	-0.0080 (0.0013)	0.0000
Household Size					0.1248 (0.0134)	0.0000	0.1249 (0.0133)	0.0000	0.1254 (0.0134)	0.0000	0.1178 (0.0144)	0.0000	0.1173 (0.0147)	0.0000	0.1172 (0.0148)	0.0000
Sex																
Female							-0.0191 (0.0364)	0.6000	-0.0189 (0.0359)	0.5990	-0.0276 (0.0355)	0.4360	-0.0239 (0.0345)	0.4870		
Male -Ref																
Education																
Educated									0.3189 (0.0435)	0.0000	0.2734 (0.0399)	0.0000	0.2103 (0.0348)	0.0000	0.2097 (0.0354)	0.0000
Not Stated – Ref																
Employment Status																
Employed											0.4029 (0.0898)	0.0000	-0.2008 (0.1123)	0.0740		
Unemployed-Ref																
Occupation																
Office Related Work (ORW)													0.8114 (0.0672)	0.0000	0.6227 (0.0779)	0.0000
Non-Office Related Work (NORW)													0.5974 (0.1071)	0.0000	0.4073 (0.0951)	0.0000
Not -Classified -Ref																
Random Component																
σ^2_{PARISH}	0.037947		0.034782		0.033452		0.033416		0.028090		0.028527		0.026634		0.027093	
σ^2_{ED}	0.077674		0.07958		0.086084		0.088084		0.079468		0.077395		0.074365		0.074365	
$\sigma^2_{HOUSEHOLD}$	0.795842		0.781986		0.705768		0.705768		0.701574		0.686247		0.678152		0.678646	
AIC	5364.57		5334.80		5151.05		5152.81		5133.63		5091.66		5068.51		5066.083	

PE=Parameter Estimates, SE = Standard Error, AIC = Akaike Information Criterion

6.8 Analysis of the CHNS -1989 &2011 Survey Data

The China Health and Nutrition Survey is an ongoing international study between the Carolina Population Centre at the University of North Carolina at Chapel Hill and the National Institute for Nutrition and Health (NINH, formerly the National Institute of Nutrition and Food Safety) at the Chinese Centre for Disease Control and Prevention (CCDC). This survey covered a range of Chinese provinces which varied substantially in geography, economic development, public resources, and health indicators.

The survey data was downloaded from the Institute's website at:
http://cpc.unc.unnc.edu/projects/china/about/proj_desc/survey.
Based on the objectives of this study, two separate data set were extracted for the years 1989 and 2011, which focused on reported income and socio-demographic variables of the survey respondents. This was to enable the modelling of Chinese population income dynamics for the years 1989 and 2011. Three thousand and fifty-two (3,052) subjects' information in the age range 16 to 65 years were extracted from 1,760 households in 161 communities spread across four strata in eight provinces in the 1989 data set. In 2011, a total of 2,109 subjects' information in the age range 16 to 65 years were extracted from 1,188 households in 167 communities spread across four strata in 12 provinces.

Table 6.6 contains the summary of the variables extracted for this analysis. The list included gender, marital status, years of schooling, reported income with imputation, and without imputation. In the summary, the data extracted for 1989 showed that the majority of individuals for the age group 16 to 65 years were male (52.6%), married (87.2%), and rural folks (69.3%) with an average of 15.31 years of schooling and average age of 35.03 years. However, in the 2011 data, the majority of individuals for the age group 16 to 65years were female (53.3%), married (92.7%), city dwellers (35.7%) with an average of 24.01 years of schooling and average age of 45.95 years.
These two datasets showed sharp contrast in the years of schooling of city dwellers and rural folks, and also the spread of the survey in 1989 and 2011 across the provinces in China. The natural logarithm of the reported income,

denoted as y_{jklm} was used in the two analyses for the years 1989 and 2011, respectively.

Table 6.8: Profile of Individuals in the Extracted Data from China Health and Nutrition Survey

	1989			2011	
CATEGORICAL VARIABLE					
VARIABLE	FREQUENCY	PERCENTAGE		FREQUENCY	PERCENTAGE
GENDER					
MALE	1606	52.6		984	46.7
FEMALE	1446	47.4		1125	53.3
TOTAL	3052	100		2109	100
MARITAL STATUS					
MARRIED	2662	87.2		1955	92.7
NOT MARRIED	390	12.8		154	7.3
TOTAL	3052	100		2109	100
PROVINCE					
BEIJING				598	28.4
LIAONING	303	9.9		7	0.3
HEILONGJIANG				20	0.9
SHANGHAI				640	30.3
JIANGSU	355	11.6		89	4.2
SHANDONG	335	11		57	2.7
HENAN	333	10.9		22	1.0
HUBEI	470	15.4		41	1.9
HUNAN	383	12.5		38	1.8
GUANGXI	478	15.7		29	1.4
GUIZHOU	395	12.9		40	1.9
CHONGQING				528	25.0
TOTAL	3052	100		2109	100
STRATUM					
CITY	173	5.7		753	35.7
SUBURBAN	449	14.7		663	31.4
TOWN OR COUNTY CAPITAL	316	10.4		433	20.5
RURAL VILLAGE	2114	69.3		260	12.3
TOTAL	3052	100		2109	100
MISSING INDICATOR					
MISSING	49	1.6		164	7.8
NOT MISSING	3003	98.4		1945	92.2
TOTAL	3052	100		2109	100
CONTINUOUS VARIABLE					
	MEAN	SD		MEAN	SD
AGE	35.03	10.14		45.95	11.94
YEARS OF SCHOOLING	15.31	8.81		24.01	7.6
REPORTED INCOME WITHOUT IMPUTATION	1371.93	1652.96		28733.88	32007
REPORTED INCOME WITH IMPUTATION	1376.01	1655.09		29383.51	32431.9

6.9 Result of the CHNS 1989 Analysis

There were 3,052 reported income cases in the 1989 dataset. Of the 3,052 cases, 49 were imputed representing 1.6% of the data. This led to two separate models: Model -IV D1 to Model -IV D8 (Table 5.7) using reported income without missing data and another denoted by Model -IVE1 to Model -IVE6 using reported income with missing data (Table 5.8). Both analyses were under the presumption of four-level multilevel model involving, Provinces at the highest level, followed by County-City level, then followed by Community and lastly the individual level. The four-level model enable estimates of the variation at each level in the overall variation in the outcome variable.

In Table 6.7, the estimated parameters for the Models: Model-IVD1 to Model-IVD6 for reported income without missing data were found to have a lower standard errors of estimate and AIC value when compared to the estimated parameters for the weighted Model-IVE1 to Model-IVE8 in Table 6.8 based on the reported income with missing data. In essence, the weight had contributed to the higher standard errors and AIC values. However, the final predictor variables are similar for the final Models in Tables 6.7 and 6.8 respectively. The significant (p-value less than 0.05) predictor variables for the reported income in the Year 1989 include age, gender and years of schooling. The noticeable difference in the Models is the higher standard errors and higher AIC values from the weighted Models arising from the adjustment for the missing data and the non-response of the individuals in the data. The ICC values showed that the majority (85.2%) of the variation in the reported income without missing data and 84.3% for the reported income with missing data existed at the individual level (see Table 6.11). This compared to the percentage of variation in the reported income at the Province, County-City and Neighbourhood -Village levels respectively.

6.10 Result of the CHNS 2011 Analysis

The CHNS 2011 extracted data had 2,109 individuals of which 164 or 7.8% did not provide the income. A series of multilevel models were investigated to identify variations in the data structure. Of all the multilevel models investigated, variations were found to exist in the three-level model starting with individual nested within the Community and nested within the province. Table 6.9 contains the parameter estimates for Model-IIIF1 to Model-IIIF6, involving income variable without missing data while Table 6.10 contains parameter estimates for Model -IIIG1 to Model -IIIG6 involving income variable with missing data. A comparable analysis revealed that standard errors in Table 6.9 are lower than the errors in Table 6.10 which due to the incorporation of the item non-response weight. Similar results were observed for the AIC values. Of all the investigated predictor variables in the formulated Models in Tables 6.9 and 6.10, gender and years of schooling were found to be significant (p-value less than 0.05). The use of the item non-response weight did not affect the final variables in either Model (Model -IIIF6, Model-IIIG6)

The ICC estimates for the 2011 dataset in Table 6.11 showed that the majority (73.97%) of the variation in the reported income without missing data and 74.86% of the variation in the reported income with missing data existed at the individual level.

Table 6.9: Parameter Estimates for Four-Level Model of Reported Income without Missing Data - Year 1989 for Subjects 16 to 65 Years

	MODEL IV-D1		MODEL IV-D2		MODEL IV-D3		MODEL IV-D4		MODEL IV-D5		MODEL IV-D6	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	**6.7190**	**0.0000**	**6.4072**	**0.0000**	**6.3134**	**0.0000**	**6.0161**	**0.0000**	**5.9052**	**0.0000**	**6.0191**	**0.0000**
	(0.0736)		**(0.1001)**		**(0.1038)**		**(0.1217)**		**(0.1362)**		**(0.1217)**	
Age			**0.0089**	**0.0000**	**0.0096**	**0.0000**	**0.0135**	**0.0000**	**0.0143**	**0.0000**	**0.0137**	**0.0000**
			(0.0019)		**(0.0019)**		**(0.0021)**		**(0.0021)**		**(0.0021)**	
Gender												
Male					**0.1324**	**0.0010**	**0.0639**	**0.1170**	**0.0572**	**0.1620**		
					(0.0381)		**(0.0407)**		**(0.0409)**			
Female-Ref												
Years of Schooling							**0.0127**	**0.0000**	**0.0122**	**0.0000**	**0.0142**	**0.0000**
							(0.0027)		**(0.0027)**		**(0.0025)**	
Marital Status												
Married									**0.1087**	**0.0720**		
									(0.0603)			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	**0.0187**		**0.0194**		**0.0194**		**0.0229**		**0.0227**		**0.0233**	
$\sigma^2_{County-City}$	**0.0529**		**0.0522**		**0.0537**		**0.0436**		**0.0435**		**0.0419**	
$\sigma^2_{Neighbourhood/Village}$	**0.1190**		**0.1182**		**0.1172**		**0.1072**		**0.1068**		**0.1065**	
$\sigma^2_{\text{Individual}}$	**1.0943**		**1.0866**		**1.0824**		**1.0777**		**1.0766**		**1.0789**	
AIC	**9128**		**9108**		**9098**		**9079**		**9078**		**9079**	

PE = Parameter Estimates, SE = Standard Error, AIC = Akaike Information Criterion

Table 6.10: Parameter Estimates for Weighted Four-Level Model of Reported Income with Missing Data - Year 1989 for Subjects 16 to 65 Years

	MODEL IV-E1		**MODEL IV-E2**		**MODEL IV-E3**		**MODEL IV-E4**		**MODEL IV-E5**		**MODEL IV-E6**	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	**6.7061**	**0.0000**	**6.3836**	**0.0000**	**6.2882**	**0.0000**	**5.9940**	**0.0000**	**5.9064**	**0.0000**	**6.0029**	**0.0000**
	(0.0815)		**(0.1175)**		**(0.1154)**		**(0.1316)**		**(0.1423)**		**(0.1283)**	
Age			**0.0092**	**0.0010**	**0.0099**	**0.0000**	**0.0137**	**0.0000**	**0.0143**	**0.0000**	**0.0139**	**0.0000**
			(0.0028)		**(0.0024)**		**(0.0030)**		**(0.0031)**		**(0.0027)**	
Gender												
Male					**0.1417**	**0.2520**	**0.0754**	**0.5760**	**0.0702**	**0.6110**		
					(0.1237)		**(0.1348)**		**(0.1334)**			
Female-Ref												
Years of Schooling							**0.0123**	**0.0000**	**0.0120**	**0.0000**	**0.0141**	**0.0000**
							(0.0031)		**(0.0031)**		**(0.0018)**	
Marital Status												
Married									**0.0856**	**0.1490**		
									(0.0593)			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	**0.0210**		**0.0214**		**0.0216**		**0.0255**		**0.0255**		**0.0261**	
$\sigma^2_{County-City}$	**0.1016**		**0.1001**		**0.1015**		**0.0884**		**0.0887**		**0.0864**	
$\sigma^2_{Neighbourhood/Village}$	**0.0794**		**0.0789**		**0.0782**		**0.0714**		**0.0711**		**0.0708**	
$\sigma^2_{\text{Individual}}$	**1.0897**		**1.0889**		**1.0841**		**1.0791**		**1.0785**		**1.0806**	
AIC	**556372**		**554971**		**554131**		**552879**		**552759**		**553087**	

PE =Parameter Estimate, SE=Standard Error, AIC = Akaike Information Criterion

Table 6.11: Parameter Estimates for Three-Level Model of Reported Income without missing data - Year 1989 for Subjects 16 to 65 Years(n=3052)

	MODEL IV-A1		**MODEL IV-B1**		**MODEL IV-C1**		**MODEL IV-D1**		**MODEL IV-E1**		**MODEL IV-F1**	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	**6.760389**	**0.0000**	**6.4368**	**0.0000**	**6.346416**	**0.0000**	**5.935446**	**0.0000**	**5.814351**	**0.0000**	**5.938543**	**0.0000**
	(0.0710017)		**(0.099857)**		**(0.1037023)**		**(0.1211603)**		**(0.1364796)**		**(0.1210903)**	
Age			**0.00917**	**0.0000**	**0.0097959**	**0.0000**	**0.015186**	**0.0000**	**0.0160107**	**0.0000**	**0.0152879**	**0.0000**
			(0.019705)		**(0.0019762)**		**(0.0021357)**		**(0.0021777)**		**(0.0021331)**	
Gender												
Male					**0.1296162**	**0.0010**	**0.0373579**	**0.3730**	**0.0301265**	**0.4740**		
					(0.0395742)		**(0.0419098)**		**(0.0420557)**			
Female-Ref												
Years of Schooling							**0.0171333**	**0.0000**	**0.0166797**	**0.0000**	**0.0179623**	**0.0000**
							(0.0026803)		**(0.0026892)**		**(0.025161)**	
Marital Status												
Married									**0.1182221**	**0.0560**		
									(0.0618875)			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	**0.0209**		**0.0222**		**0.0227**		**0.0251**		**0.1571**		**0.0252**	
$\sigma^2_{Neighbourhood/Village}$	**0.0627**		**0.0605**		**0.0603**		**0.0482**		**0.2193**		**0.0478**	
$\sigma^2_{\text{Individual}}$	**1.1885**		**1.1803**		**1.1761**		**1.1622**		**1.0774**		**1.1626**	
AIC	**9257.8**		**9238.2**		**9229.5**		**9191.2**		**9189.57**		**9190**	

PE =Parameter Estimate, SE=Standard Error, AIC = Akaike Information Criterion

Table 6.12: Parameter Estimates for Weighted Three-Level Model of Reported Income without Imputed data - Year 1989 for Subjects 16 to 65 Years

	MODEL IV-A2		**MODEL IV-B2**		**MODEL IV-C2**		**MODEL IV-D2**		**MODEL IV-E2**		**MODEL IV-F2**	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	**6.756339**	**0.0000**	**6.39597**	**0.0010**	**6.301246**	**0.0000**	**5.869976**	**0.0000**	**5.7665**	**0.0000**	**5.9039**	**0.0000**
	(0.077009)		**(0.13054)**		**(0.129399)**		**(0.1516843)**		**(0.1645051)**		**(0.144765)**	
Age			**0.0102184**	**0.0000**	**0.0108658**	**0.0000**	**0.0163964**	**0.0000**	**0.0171027**	**0.0000**	**0.0158784**	**0.0000**
			(0.00313)		**(0.0028902)**		**(0.0032)**		**(0.0031)**		**(0.0030157)**	
Gender												
Male					**0.134377**	**0.2540**	**0.0420**	**0.7310**	**0.0359656**	**0.7650**		
					(0.1178)		**(0.12203)**		**(0.1202)**			
Female-Ref												
Years of Schooling							**0.0181041**	**0.0000**	**0.017725**	**0.0000**	**0.0185**	**0.0000**
							(0.0026086)		**(0.00266)**		**0.0023973**	
Marital Status												
Married									**0.1005249**	**0.0980**		
									0.0607725			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	**0.0414**		**0.0420**		**0.0425**		**0.0423**		**0.0419**		**0.0423**	
$\sigma^2_{\textit{Neighbourhood/Village}}$	**0.0412**		**0.0398**		**0.0396**		**0.0310**		**0.0308**		**0.0297**	
$\sigma^2_{\text{Individual}}$	**1.1961**		**1.1860**		**1.1816**		**1.1667**		**1.1657**		**1.1694**	
AIC	**574339.9**		**572696.5**		**571989**		**569257.3**		**569097.7**		**563146.6**	

PE =Parameter Estimate, SE=Standard Error, AIC = Akaike Information Criterion

Table 6.13: Parameter Estimates for Three-Level Model of Reported Income without Missing Data - Year 2011 for Subjects 16 to 65 Years

	MODEL III-F1		MODEL III-F2		MODEL III-F3		MODEL III-F4		MODEL III-F5		MODEL III-F6	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	**9.6737**	**0.0000**	**10.1311**	**0.0000**	**10.0296**	**0.0000**	**8.4630**	**0.0000**	**8.3943**	**0.0000**	**8.5067**	**0.0000**
	0.1336		**0.1602**		**0.1574**		**0.1631**		**0.1742**		**0.1281**	
Age			**-0.0108**	**0.0000**	**-0.0113**	**0.6680**	**0.0007**	**0.6680**	**0.0006**	**0.7100**		
			0.0016		**0.0016**		**0.0017**		**0.0017**			
Gender												
Male					**0.2750**	**0.0000**	**0.1802**	**0.0000**	**0.1765**	**0.0000**	**0.1819**	**0.0000**
					0.0369		**0.0355**		**0.0357**		**0.0353**	
Female-Ref												
Years of Schooling							**0.0469**	**0.0000**	**0.0470**		**0.0463**	**0.0000**
							0.0030		**0.0030**		**0.0027**	
Marital Status												
Married									**0.0768**	**0.2560**		
									0.0676			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	**0.1779**		**0.2124**		**0.2016**		**0.1224**		**0.1235**		**0.1249**	
$\sigma^2_{Neighbourhood/Village}$	**0.0569**		**0.0592**		**0.0592**		**0.0273**		**0.0272**		**0.0275**	
$\sigma^2_{\text{Individual}}$	**0.7397**		**0.7238**		**0.7051**		**0.6367**		**0.6363**		**0.6366**	
AIC	**5428**		**5387**		**5334**		**5105**		**5106**		**5103**	

PE = Parameter Estimates, SE = Standard Error, AIC = Akaike Information Criterion

Table 6.14: Parameter Estimates for Weighted Three-Level Model of Reported Income with Missing Data - Year 2011 for Subjects 16 to 65 Years

	MODEL III-G1		MODEL III-G2		MODEL III-G3		MODEL III-G4		MODEL III-G5		MODEL III-G6	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	**9.6557**	**0.0000**	**10.09018**	**0.0000**	**9.98304**	**0.0000**	**8.45406**	**0.0000**	**8.394639**	**0.0000**	**8.556097**	0.0000
	(0.1437)		**(0.243043)**		**(0.23154)**		**(0.2239)**		**(0.240086)**		**(0.138493)**	
Age			**-0.01019**	**0.0030**	**-0.0109**	**0.0010**	**0.00096**	**0.7590**	**0.000904**	**0.7790**		
			(0.003384)		**(0.00321)**		**(0.00313)**		**(0.003221)**			
Gender												
Male					**0.2903**		**0.19322**		**0.190192**	**0.0000**		
					(0.02499)		**(0.03829)**		**(0.03477)**			
Female-Ref												
Years of Schooling							**0.04555**	**0.0000**	**0.045628**	**0.0000**	**0.046806**	0.0000
							(0.00515)		**(0.00525)**		**(0.00554)**	
Marital Status												
Married									**0.065247**	**0.466**		
									(0.08957)			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	**0.2241**		**0.2595**		**0.2470**		**0.1619**		**0.1626**		**0.1719**	
$\sigma^2_{\textit{Neighbourhood/Village}}$	**0.0434**		**0.0441**		**0.0443**		**0.0201**		**0.0201**		**0.0196**	
$\sigma^2_{\text{Individual}}$	**0.7599**		**0.7463**		**0.7253**		**0.6579**		**0.6576**		**0.6675**	
AIC	**65776.46**		**65329.7**		**64615.51**		**61960.92**		**61952.49**		**62316.8**	

PE =Parameter Estimate, SE=Standard Error, AIC = Akaike Information Criterion

Table 6.15: Parameter Estimates for Four-Level Model of Reported Income without Missing Data - Year 2011 for Subjects 16 to 65 Years

	MODEL IV-A3		**MODEL IV-B3**		**MODEL IV-C3**		**MODEL IV-D3**		**MODEL IV-E3**		**MODEL IV-F3**	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	9.6629	0.0000	10.079	0.0000	9.9859	0.0000	8.5576	0.0000	8.4973	0.0000	8.5843	0.0000
	0.1390		0.1646		0.1619		0.1674		0.1777		0.1319	
Age			- 0.0099	0.0000	- 0.0105	0.0000	0.0004	0.7970	0.0003	0.8430		
			0.0016		0.0016		0.0017		0.0017			
Gender												
Male					0.2699	0.0000	0.1846	0.0000	0.1814	0.0000	0.1857	0.0000
					0.0356		0.0348		0.0349		0.0346	
Female-Ref												
Years of Schooling							0.0429	0.0000	0.0429	0.0000	0.0425	0.0000
							0.0348		0.0031		0.0027	
Marital Status									0.0683	0.3070		
Married									0.0667			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	0.1981		0.2330		0.2223		0.1321		0.1332		0.1337	
$\sigma^2_{County-City}$	0.0000		0.0066		0.0079		0.0056		0.0055		0.0058	
$\sigma^2_{Neighbourhood/Village}$	0.1142		0.1110		0.1103		0.0591		0.0590		0.0593	
$\sigma^2_{\text{Individual}}$	0.6808		0.6677		0.6495		0.6058		0.6054		0.6056	
AIC	5334		5299		5244		5064		5065		5062	

Table 6.16: Parameter Estimates for Weighted Four-Level Model of Reported Income with Missing Data - Year 2011 for Subjects 16 to 65 Years

	MODEL IV-A4		**MODEL IV-B4**		**MODEL IV-C4**		**MODEL IV-D4**		**MODEL IV-E4**		**MODEL IV-F4**	
Variable	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**	**PE (SE)**	**p-value**
Fixed Effects												
Constant	9.5715	0.0000	9.9768	0.0000	9.8886	0.0000	8.4479	0.0000	8.3814	0.0000	8.4461	0.0000
	0.1669		0.25959		0.2474		0.2215		0.2309		0.1526	
Age			-0.0099	0.0070	-0.0111	0.0020	-0.0000	0.9930	-0.0001	0.9740	0.2066	0.0000
			0.0037		0.0036		0.0034		0.0035		0.0356	
Gender												
Male					0.2949	0.0000	0.2067	0.0000	0.20311	0.0000	0.0433	0.0000
					0.0241		0.0363		0.0328		0.0052	
Female-Ref												
Years of Schooling							0.0433	0.0000	0.0434	0.0000		
							0.0047		0.0049			
Marital Status									0.0743	0.3300		
Married									0.0763			
Never married-Ref												
Random Component												
$\sigma^2_{\text{Province}}$	0.2296		0.2330		0.2223		0.1321		0.1332		0.1337	
$\sigma^2_{County-City}$	0.2604		0.0066		0.0079		0.0056		0.0055		0.0058	
$\sigma^2_{Neighbourhood/Village}$	0.0738		0.1110		0.1103		0.0591		0.0590		0.0593	
$\sigma^2_{\text{Individual}}$	0.6808		0.6677		0.6495		0.6058		0.6054		0.6056	
AIC	5334		5299		5244		5064		5065		5062	

Table 6.17: Variance and Intraclass Estimates for the CHNS Data-1989 and 2011

Four Level Model	**1989**
$y_{jklm} = x_{jklm}^{T}\beta + u_j + v_{jk} + \tau_{jkl} + \varepsilon_{jklm}$ Where y_{jklm} represents reported individual income without missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, $\sigma^2_{\tau_{jkl}}$, $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, County-City, Neighbourhood /Village and individual levels respectively.	σ^2_u = 0.0187 ρ_u = 0.015 σ^2_v = 0.0529 ρ_v = 0.041 σ^2_τ = 0.1190 ρ_τ = 0.093 σ^2_ε = 1.0943 ρ_ε = 0.852
$y_{jklm} = x_{jklm}^{T}\beta + u_j + v_{jk} + \tau_{jkl} + \varepsilon_{jklm}$ Where y_{jklm} represents reported individual income without imputed data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, $\sigma^2_{\tau_{jkl}}$, $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, County-City, Neighbourhood / Village and individual levels respectively.	σ^2_u = 0.0210 ρ_u = 0.016 σ^2_v = 0.1016 ρ_v = 0.078 σ^2_τ = 0.0794 ρ_τ = 0.061 σ^2_ε = 1.0971 ρ_ε = 0.845
Three Level Model	**1989**
$y_{jkl} = x_{jkl}^{T}\beta + u_j + v_{jk} + \varepsilon_{jkl}$ Where y_{jkl} represents reported individual income without missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, , $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, Neighbourhood/Village and individual levels respectively.	σ^2_u = 0.021 ρ_u = 0.0164 σ^2_v = 0.063 ρ_v = 0.04923 σ^2_ε = 1.189 ρ_ε = 0.93429
$y_{jkl} = x_{jkl}^{T}\beta + u_j + v_{jk} + \varepsilon_{jkl}$ Where y_{jkl} represents reported individual income with missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, , $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, Neighbourhood-Village and individual levels respectively.	σ^2_u = 0.0414 ρ_u = 0.0323 σ^2_v = 0.0412 ρ_v = 0.0322 σ^2_ε = 1.1961 ρ_ε = 0.9354

Table 6.18: Variance and Intraclass Estimates for the CHNS Data-2011

Four Level Model	**2011**
$y_{jklm} = x^{T}_{jklm}\beta + u_j + v_{jk} + \tau_{jkl} + \varepsilon_{jklm}$ Where y_{jklm} represents reported individual income without missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, $\sigma^2_{\tau_{jkl}}$, $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, County-City, Neighbourhood/Village and individual levels respectively.	σ^2_u = 0.1981 ρ_u = 0.199 σ^2_v = 0.0000 ρ_v = 0.000 σ^2_τ = 0.1142 ρ_τ = 0.115 σ^2_ε = 0.6808 ρ_ε = 0.686
$y_{jklm} = x^{T}_{jklm}\beta + u_j + v_{jk} + \tau_{jkl} + \varepsilon_{jklm}$ Where y_{jklm} represents reported individual income with missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, $\sigma^2_{\tau_{jkl}}$, $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, County-City, Neighbourhood/Village and individual levels respectively.	σ^2_u = 0.2296 ρ_u = 0.184 σ^2_v = 0.2604 ρ_v = 0.209 σ^2_τ = 0.0738 ρ_τ = 0.059 σ^2_ε = 0.6808 ρ_ε = 0.547
Three Level Model	**2011**
$y_{jkl} = x^{T}_{jkl}\beta + u_j + v_{jk} + \varepsilon_{jkl}$ Where y_{jkl} represents reported individual income without missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, , $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, Neighbourhood/Village and individual levels respectively.	σ^2_u = 0.1779 ρ_u = 0.183 σ^2_v = 0.0569 ρ_v = 0.058 σ^2_ε = 0.7397 ρ_ε = 0.759
$y_{jkl} = x^{T}_{jkl}\beta + u_j + v_{jk} + \varepsilon_{jkl}$ Where y_{jkl} represents reported individual income with missing data while $\sigma^2_{u_j}$, $\sigma^2_{v_{jk}}$, , $\sigma^2_{\varepsilon_{jkl}}$ represents the variances at the Province, Neighbourhood -Village and individual levels respectively.	σ^2_u = 0.2242 ρ_u = 0.223 σ^2_v = 0.0392 ρ_v = 0.039 σ^2_ε = 0.7405 ρ_ε = 0.738

6.11 Summary of the Analyses

Two separate datasets from the Jamaican and Chinese population were analysed with emphasis of household expenditure of meals purchased away from home and income dynamics, respectively. In the JSCL 2007 analysis, the result provided the evidence that type of weighted models was associated with the proportion of variation in the data hierarchy. The best model had the least AIC and also incorporates all the hierarchy in the survey design.

The findings from the analysis of the China Health Survey Data showed that standard error of estimates is usually lower for parameter estimates from the weighted model than unweighted model especially in this scenario where some of the data values are imputed. Similarly, depending on the profile of the socio-demographics of the survey respondents, there will be varying predictors for probability of response and ultimately the item non-response weight, the inverse of the response probability. This has led to different predictors in the logistics models for years 1989 and 2011. The result also showed that depending on the level of significance, several models can be created.

In summary, the advantage of the item non- response weighted multilevel model was reduced standard error of estimate. In addition, conducting a multilevel model was a right approach to analyse survey data from non-uniform sampling method.

Chapter 7

Discussion, Conclusion and Limitation

7.1 Discussion

The first and second objectives of the study were developed to identify the weighing adjustment method that produces the least bias parameter estimate when addressing the unequal probability of selection and missing data problems in multilevel models. In addition to the least bias and variance, the objectives also included the conditions which are favourable for this least bias estimate in a multilevel modelling.

The findings of the simulation study on the parameter estimates from the multilevel model suggested that reliable estimates are possible when weights are applied to address the unequal probability of selection at each level of the sampling design. This is comparable to the suggestions from DuMouchel and Duncan (1983) as well as Pfefferman (1993) regarding the use of weighting adjustment methods to address the bias from unequal selection probability in regression models involving stratified samples.

Reliable estimates were also observed when adjusting for item non response because the missing data in the sample confirmed the usefulness of item non response weighting adjustments as expressed in Skinner and Arrigo (2011); Bethlehem et.al (2011); Barbara and Stephen(2001). However, sampling weights adjustments for unequal probability of selection of sample units with further adjustments for item non response produced the least bias estimate from the simulation study. This suggests that better estimates can be produced when both unequal probability of sample selection at each level of the stratification along while also addressing the missing data in the sample. This concept is not common in the literature but serves as a way forward for researchers having this type of problems.

As expected, the findings revealed that the less missing data, the better the estimates. The weighting adjustment methods perform better when the missing data is 20% or less under the Missing at Random (MAR) and also Missing not at Random mechanisms. Performance of the weighting adjustment methods decreases linearly as the percentage of missing increases from 40% to 60%. In addition, the study findings also revealed that adjusting for missing data and unequal sample selection in survey sampling design simultaneously during the estimation of model parameters produces better estimate and performance of the weight adjustment methods.

The earlier findings were only in relation to a fixed intraclass correlation coefficient (ICC). The inclusion of varying ICC values was to address the third study objective of the reliability of estimates from weighted multilevel model under varying ICC values. The findings revealed that at higher ICC values the parameter estimates from the weighting adjustment methods produces smaller root mean square error (RMSE) while at lower ICC values higher root mean square error were produced. This simply implies that the weighting adjustments methods performs better at higher ICC. The ICC values in an indication variation within and between the cluster based on the sample size. This finding is synonymous to the issues in the literature regarding cluster size in multilevel models. The varying sizes of clusters thus influence the variation between and withing clusters and by extension the intraclass correlation coefficient (Hox, 1998,2010; Snijders and Bosker, 2012).

For the fourth objective of the thesis, which seeks to identify any possible effects of the cluster sizes at level 1 or 2 in a weighted or unweighted multilevel model on the predictor variables in the model, the findings clearly shows that there is no effect.

7.2 Conclusions

From the study, the simulations revealed adjusting for unequal probability of selection of units in a multilevel model is necessary more importantly simultaneous adjustments for missing data when the information for such adjustment are available. Researchers conducting multilevel model should be free to apply the method to account for clustering so far the value of the ICC is not zero and the clustering was part of the data design , that was integral in the design information of the survey. Reliable parameter estimates are generally guaranteed when the proportion of missing data in the study variable is 20% or less and weights in the parameter estimation should be incorporated when the information is available.

7.3 Limitation

The simulation findings are only limited to one hundred repetitions due to computer resources and time factor to complete the study. Lack of information on the nonrespondents who constitute the unit-non response make it impossible to explore adjustment for unit non-response. This is the reality which many secondary data users are usually faced around the world and did not provide the opportunity for comparison with the other methods.

7.4 Future Research

The future research from this study would be comparative analysis of multiple imputation with item non response to determine the effects of both on parameters from weighted multilevel linear model. This will provide guidance on which conditions support weighting adjustment and which conditions different the use of imputation from weighting adjustments.

Bibliography

Allison, P.D (2000) . Missing data. Thousand Oaks , CA: Sage Publications.

Andridge and Little (2010). A review of hot deck imputation for survey non-response. International statistical review, 2010 - Wiley Online Library

Asparouhov, Tihomir (2006) General Multi-Level Modeling with Sampling Weights, Communications in Statistics - Theory and Methods, 35:3, 439-460, DOI: 10.1080/03610920500476598

Bai, J., Wahl, T. I., Lohmar, B. T., & Huang, J. (2010). Food away from home in Beijing: Effects of wealth, time and free meals. *China Economic Review*, *21*(3), 432.

Bates, D. M., and Pinherio, J. C. (1998). Computational Methods for Multilevel Modeling. Retrieved from: https://www.researchgate.net/publication/2753537_Computational_Methods_for Multilevel_Modelling

Bethlehem J.,Cobben F.,Schouten B. (2011). Handbook of Nonresponse in Household Surveys. John Wiley & Sons.Inc Hoboken, New Jersey

Brick , J.M (2013). Unit Nonresponse and Weighting Adjustments: A critical Review. Journal of Official Statistics, Vol. 29, No.3, 2013, pp 329 – 353. DOI : 10.2478/jos-2013-0026.

Byrne, P. J., Capps, O., & Saha, A. (1996). Analysis of food-away-from-home expenditure patterns for U.S. households, 1982-89. *American Journal of Agricultural Economics, 78*(3), 614-627.

Buuren S. V(2012).Flexible Imputation of Missing Data . CRC Press. Taylor & Francis Group . ISBN 978-1-4398-6824-9

Carle Adam C. (2009). Fitting multilevel models in complex survey data with design weights; Recommendations. BMC Medical Research Methodology. http://www.biomedcentral.com/147-2288/9/49

Carroll R.. J, Jeff Wu C.F, Ruppet D. (2014). The Effect of Estimating Weights in Weighted Least Squares. Journal of the American Statistical Association. http://amstat.tandfonline.com/loi/uasa20. Volume 83, 1988 - Issue 404

Chiou, J. M and Muller H.G (2005). Estimated estimating equations: semiparametric inference for clustered and longitudinal data. Journal of the Royal Statistical Society 2005 - Wiley Online Library **.** Volume67, Issue4 September 2005

Clarke, P. (2007). When can group level clustering be ignored? Multilevel models versus single – level models with sparse data. Theory and Methods. Journal of Epidemiology & Community Health, 2008 - jech.bmj.com 62 665-665 Published. http://dx.doi.org/10.1136/jech.2007.060798 Online First: 11 Jul 2008

Cochran, W. G 1942 Sampling theory when the sampling-units are of unequal sizes. Journal of the American Statistical Association. - amstat.tandfonline.com Journal of the American Statistical Association Volume 37, 1942 - Issue 218

Demitras, H Freels S.A, and Yucel R.M (2008). Journal of Statistical Computation and Simulation 78(1):69-84

Dey, E.L (1997). WORKING WITH LOW SURVEY RESPONSE RATES: The Efficacy of Weighting Adjustments. Research in Higher Education, Vol. 38, No. 2, 1997

Doran, H. and Bates, D., Blise, P and Dowling, M. (2007). Estimating the Multilevel Rasch Model: With the lme-4 Package. Journal of Statistical Software. April 2007, Volume 20, Issue 2. http://www.jstatsoft.org/

Dumouchel W.H & Duncan G.J (1983). Using sample survey weights in multiple regression analyses of stratified samples. Journal of the American Statistical 1983 - Taylor & Francis. Journal of the American Statistical Association Volume 78, 1983 - Issue 383

Durrant, Gabriele B. and Steele, Fiona (2009) Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. Journal of the Royal Statistical Society: series A (statistics in society), 172 (2). pp. 361-381. ISSN 0964-1998
DOI: 10.1111/j.1467-985X.2008.00565.x

Fagbamigbe A.F and Bakre B. B (2018). Evaluating Likelihood Estimation Methods in Multilevel Analysis of Clustered Survey Data. African Journal of Applied Statistics. Vol. 5 (1).2018, pages 352-376.

Farell, S. and Ludwig C. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. Psychon Bull Rev.2008 December; 15(6): 1209 -1217. Doi:10.3758/PBR.15.6.1209

Feder, M., Nathan, G., and Pfeffermann D., (2000). Multilevel Modeling of Complex Longitudinal Data With Tme Varying Random Effects. Survey Methodology June 2000. Vol.26. No. 1 pp.53 -65. Statistics Canada.

Finch, W. H., Bolin J. E., and Kelley K. (2014). *Multilevel modeling using R*. CRC Press.

Gardiner, C, J., Luo, .Z, and Roman, L.A. (2009). Fixed effects, random effects and GEE: What are the differences? Statistics in Medicine. Statist. Med. 2009; 28: 221 – 239. Wiley Inter Science. www.interscience.wiley.com . DOI: 10.1002/sim.3478.

Gibbons, R. D., and Hedeker, D. (1997). Random Effects Probit and Logistic Regression Models for Three-Level Data. *Biometrics, 53*(4), 1527-37.

Goldstein, .H (1995).Hierarchical Data Modeling in the Social Science. Journal of Educational and Behavioural Statistics Summar 1995. Vol.20 Number 2.

Goldstein, .H (1986). Multilevel Linear Model Analysis Using Iterative Generalised Least Squares. Biometrika 73 (1) 43 -56.

Goldstein, .H (1995). Multilevel Statistical Models (Second Edition). London: Edward Arnold; New York: Halsted Press.

Gourieroux C., Monfort A and Trognon A (1984). Pseudo Maximum Likelihood Methods: Theory. Econometrics. May 1984, Vol 52. No. 3 pp.681 – 700.

Grace Y, Yi, Rao J.N.K and Li ,H. (2016). A weighted composite likelihood approach for analysis of survey data under two – level models. Statistica-Sinica 26 (2016), 569 – 587. http://dx.doi.org/10.5705/ss.2013.383

Graubard, B.I and Korn, E.L . (1995). Weighted and Unweighted Estimates from a Sample Survey. The American Statistician. Vol 49 , No.3 (Aug., 1995), pp.291-295. https://www.jstor.org/stable/2684203

Graubard, B.I and Korn, E.L . (1996). Modeling the sampling design in the analysis of health surveys. Statist. Meth. Med. Res., 5. 263 -281.

Griebel M , Heiss F, Oettershagen J, Weiser C (2019). Maximum Approximated Likelihood Estimation. https://arxiv.org/pdf/1908.04110.pdf

Grilli, L. and Pratesi, M.(2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. Surv. Methodol, **3**0**,** 93-103

Groves, .R (2006). Nonresponse Rates and Nonrespose Bias in Household Surveys. Public Opinion Quarterly 70(5):646 · January 2006. DOI: 10.1093/poq/nfl033

Grund, S, Ludtke, O., & Robitzsch A . (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. Behavior Research Methods, 48, 640 – 649. DOI:10.3758/s13428-015-0590-3

Grund, S, Ludtke, O., & Robitzsch A . (2016). Multiple imputation of missing data for multilevel models: Simulations and Recommendations Organizational Research Methods 2018, Vol. 21(1) 111-149. Sagepub.com

Gustavsson, S.M, Johannesson S, Sallsten G, and Anderson E.M (2012). Linear Maximum Likelihood Regression Analysis for Untransformed Log - Normally Distributed Data. Open Journal of Statistics, 2012, 2, 389 – 400. http://ds.doi.org/10.4236/ois.2012.24047

Hansen M.H and Hurwitz W.N (1934). On the theory of sampling from finite populations. Annals of Mathematics and Statistics, 14, 333-362.

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z., and Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine, 20,* 15, 2219-2241. https://doi:10.1002/sim918

Himelein (2013). Weight calculations for panel surveys with sub-sampling and split-off tracking. elibrary.worldbank.org.
https://doi.org/10.1596/1813-9450-6373

Hirotugu, A. (1978). A Bayesian Analysis of the Minimum AIC procedure. *Annals of the Institute of Statistics and Mathematics/Edited by the Institute of Statistical Mathematics, 30*(1), Part A, 9-14.

Holt, D. and Elliot, D. (1991). Methods of Weighting for Unit Non-Response. The Statistician, Vol. 40, No. 3. Special Issue: Survey Design, Methodology and Analysis (2) (1991), 333-348.

Holt, D and Smith T.M.F (1979). Post Stratification , Journal of Royal Statistical Society A, 142 , 33 – 46.

Horvitz D. G and Thompson D.J (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 1952 - amstat.tandfonline.com

Huang, F.L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*(3), 492-499. https://doi.org/10.1037/spq0000272

Iannacchione, V.G, Milne, J.G and Folson, R.E (1991). Response Probability Weight Adjustments Using Logistic Regression. Proceeding of the American Statistical Association. Section on Survey Research Methods 637-642.Institute for Social Research, University of Michigan, USA

Johnson R. David and Elliot A. Lisa (1998). Sampling Design Effects: Do They Affect the Analyses of Data from the National Survey of Families and Households?. Journal of Marriage and Family, Nov., 1998, Volume 60, No. 4 (Nov., 1998), pp.993 -1001

Kalton, G., and Flores-Cervantes, I (2003). Weighting Methods. Journal of Official Statistics 19.pp81-97.

Karl, A. T, Yang, .Y and Lohr, S.L (2014). Computation of maximum likelihood estimates for multiresponse generalized linear models with non-nested , correlated random effects. Computational Statistics and Data Analysis. www.elsevier.com/locate/csda.

Kish L and Frankel M.R (1974). Inference from complex samples. Journal of the Royal Statistical Society 1974 - Wiley Online Library

Volume36, Issue1 September 1974 Pages 1-22

Kish, L. (1965). Survey Sampling . London: Wiley.

Korn & Graubard (1995). Analysis of large health surveys: accounting for the sampling design. Journal of the Royal Statistical Society

Volume158, Issue2 1995 Pages 263-295

Korn & Graubard (1995). Examples of differing weighted and unweighted estimates from a sample survey. The American Statistician Volume 49, 1995 - Issue 3

Kovacevic M .S, and Rai S .N (2003). A Pseudo Maximum Likelihood Approach to Multilevel Modelling of Survey Data. Communication in Statistics – Theory and Methods . ISSN:0361-0926. https://www.tandfonline.com/loi/lsta20

Laird, N.M and Ware, J.H (1982). Random -Effects Models for Longitudinal Data Biometrics, Vol. 38, No.4. (Dec., 1982), pp. 963 – 974

Lee, V.E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125-41. https://doi.org/10.1207/S15326985EP3502_6

Li., X and Hedeker, D. (2012). A Three-Level -Mixed-Effects Location Scale Model With An Application To Ecological Momentary Assessment (EMA) Data. Stat. Med, 2012 November 20;31(26): 3192 -3210. Doi:10.1002/sim.5393

Lindstrom M. J and Bates D.M (1988). Newton-Raphson and EM Algorithm for Linear Mixed -Effects Models. Journal of the American Statistical Association, Volume 83, Issue 404 (Dec., 1988)

Little R.J.A and Rubin, D.B.(1987, 2002). STATISTICAL ANALYSIS WITH MISSING DATA (2nd Edition). Wiley Series in Probability and Statistics. Wiley

Ludtke, O., Robitzsch, A., & Grund , S. (2017). Multiple imputation of missing data in multilevel designs : A comparison of different strategies. Psychological Methods, 22, 141 – 165. DOI: 10.1037/met0000096

Lohr S.L(2010). Sampling: Design and Analysis, Second Edition. Brooks /Cole CENGAGE Learning.

Ma, H., Huang, J., Fuller, F., & Rozelle, S. (2006). Getting rich and eating out: consumption of food away from home in urban China. *Canadian Journal of Agricultural Economics/revue Canadienne D'agroeconomie, 54*(1), 101-119.

Martin, J. and Matheson, J. (1999) Responses to Declining Response Rates on Government Surveys, Survey Methodology Bulletin, 45, 33-37.

Madow WG and Madow, LH (1944) On the theory of systematic sampling, The Annals of Mathematical Statistics,
Vol. 15, No. 1 (Mar., 1944), pp. 1-24 (24 pages)

Madow, W.G , Olkin, L ., and Rubin ,D.B., Editors (1983). Incomplete Data in sampling survey Volume 2. Academic Press, New York

Manrique, J., and Jensen, H.H. (1998). Spanish household demand for seafood products. Economic Presentations, Posters and Proceedings. 16. http://lib.dr.iastate.edu/econ_las_conf/16

McCracken, V. A., & Brandt, J. A. (1987). Household consumption of food-away-from-home: total expenditure and by type of food facility. *American Journal of Agricultural Economics*, *69*(2), 274-284.

McNeish D.M and Stapleton L.M (2014). The Effect of Small Sample Size on Two -Level Model Estimates: A Review and Illustration . Educ Psychol Rev. DOI 10.1007/s10648-014-9287-x

Mistler, S.A (2015). A SAS macro for applying multiple imputation to multilevel data . In proceedings of the SAS Global Forum . Retrieved from http:// support sas

Nayman (1934). Method of Stratified **Sampling** and the Method of Purposive Selection, Journal of the Royal Statistical Society, 97: 558- 606 (1934)

National Research Council 2013. Nonresponse in Social Science Surveys: A Research Agenda. Washington, DC: The National Academies Press. https://doi.org/10.17226/18293

Nelder J. A., and Lee, Y. (1992). Likelihood, Quasi-Likelihood and Pseudolikelihood: Some Comparisons. http://www.jstor.com/stable/2345963

Nezlek, J.B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*(2), 842 – 60. https://doi.org/10.1111/j.1751-9004.2007. 00059.x

Oguz-Alper, .M, and Berger, Y .G (2020). Modelling multilevel data under Complex sampling designs: An empirical likelihood approach. Computational Statistics and Data Analysis. www.elsevier.com/locate/csda.

Patrician A., Patricia (2002). Focus on Research Methods Multiple Imputation for Missing Data. Research in Nursing & Health, 2002, 25 76 – 84. DOI 10.1002/nur.10015

Park ,S. and Lake, T.E, (2005). Multilevel Modeling of a Clustered Continuous Outcome: Nurses' Work Hours and Burnout. Nurs Res. 2005: 54(6): 406 - 4133

Parzen M., Lipstiz S. R., Ibrahim J. G. and Lipshultz S. (2002). A weighted estimating equation for linear regression with missing covariate data. Statistics in Medicine 2002; 21: 2421 – 2436 . DOI:10.1002/sim.1195

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale De Statistique, 61*(2), 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research, 5(*3), 239-261.

Pfeffermann, D., Moura, F. A. D. S.and Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling. *Biometrika, 93*(4), 943-959.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H.and Rasbash, J. (January 01, 1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society. Series B (statistical Methodology), 60*(1), 23-40.

Pike, G. (2007). Using Weighting Adjustments to Compensate for Survey Nonresponse. Res High Educ (2008) 49:153 – 171. Doi: 10.1007/s11162-007-9069-0

Pinherio, J.C and Bates, D.M (2000). Mixed-Effects Models in S and S-PLUS Springer -Verlag New York, Inc

Rabe-Hesketh, S., amd Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal-Royal Statistical Society Series A, 169*(4), 805-827.

Rasbash J. and Goldstein H. (1994). Efficient Analysis of Mixed Hierarchical and Cross-Classified Random Structures Using a Multilevel Model. Journal of Educational and Behavioral Statistics, Vol. 19, No.4 (Winter, 1994), pp.337-350

Raudenbush, S. and Bryk, A.S, (1986). A Hierarchical Model for Studying School Effects. Sociology of Education, Vol 59. No. 1(Jan. 1986) pp 1-17. American Sociological Association. DOI:10.2307/2112482. https://www.jstor.org/stable/2112482.

Reilly M., & Pepe M., (1997). The relationship between hot -deck multiple imputation and weighted likelihood statistics in medicine, 16 , 5-19.

Robinson, K., and Pevalin, D.J. (2016). *Multilevel modeling in plain language*. Sage.

Schafer J, Yucel R. Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics. 2002; 11:421–442.

Schafer, J.L & Schenker, N (2000). Inference with imputed cpnditional means. Journal of the American Statistical Association, 449, 144 -154

Scheaffer L Richard, Mendenhall William III and Lyman R Ott (1996). Elementary Survey Sampling 5th Edition. Duxbury Pres. ISBN 0-534 -24341-8.

Searle, S.R., Casella, G and McCulloch, C.E.(1992). Variance Components. London: Wiley.

Steeh, C., Kirgis, N., Cannon, B. and DeWitt, J. (2001) Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century, Journal of Official Statistics, 17, 227-247.

Skinner and Wakefield (2017). Introduction to the design and analysis of complex survey data. Statistical Science, 2017 - projecteuclid.org. https://projecteuclid.org/euclid.ss/1494489809

Smith T.M.F (1991). Post -Stratification. The Statistician, 40, 315 - 23

Spilker M. E and Vicini P (2002). An Evaluation of Extended vs Weighted Least Squares for Parameter Estimation in Physiological Modeling. Journal of Biomedical Information 34, 348 -364.

Singer Eleanor (2006). NONRESPONSE BIAS IN HOUSEHOLD SURVEYS. Public Opinion Quarterly, Vol 70, No.5, Special Issue 2006.pp637-645

Tabachnick, B.G., & Fidell, L.S (2000) . Using multivariate statistics (4thed). New York : Harper Collins College Publishers

Quartagno ,M, & Carpenter J.R (2016).jomo: A package for multilevel joint modeling multiple imputation (version 2.1-2)

United Nations (2005). Designing Household Survey Samples. New York 2005. Unstats.un.org/unsd/demographic/sources/surveys/Hnadbook23June05.pdf

Valliant, R., Dever, J.A, and Kreuter, F. (2013). Practical Tools for Designing and Weighting Survey Samples. Statistics for Social and Behavioural Sciences. Springer. http://www.springer.com/3463

Veiga, A., Smith, P. W. F., & Brown, J. J. (2014). The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 63*(1), 65-84.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228-243. https://doi:10.1037/a0027127

Zhang X , Holt J.B, Hua L, Wheaton, A.G , Ford E.S, Greenlund K J and Croft J.B (2014). Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System. American Journal of Epidemiology. Vol. 179, No. 8 DOI: 10.1093/aje/kwu018 Advance Access publication: March 4, 2014

Zhou, H, Chen, J. Rissanen T.H, Korrick, S.A, Hu, H, Salonen, J. T, and Longnecker, M.P (2007). An efficient sampling and inference procedure for studies with a continuous outcome. Epidemiology, 2007 July: 18(4): 461-468.doi:10.10097/EDE.0b013e31806462d3

Appendix A

Simulation Codes

```
          setwd("C:/Users/KOCI 2/Desktop/Y20AGE_SEPT_5_19")

X1=read.csv(file="exampleM0Y20.5002AGE.csv")
X1.2=read.csv(file="exampleM0Y20.1002AGE.csv")
M1=matrix(0, ncol=5,nrow=100)
M1=data.frame(M1)

for(i in 1:50)
{
  M1[i,]=X1[c(3,5,8,11,14),(i+1)]
  M1[(50+i),]=X1.2[c(3,5,8,11,14),(i+1)]
}
colnames(M1)=c("Age","Constant","log U","log V","log E")

write.csv(M1,file="M0Y20AGE.csv")

Y0=read.csv("M0Y20AGE.csv")
#
#
X1=read.csv(file="exampleM1Y20.5002AGE.csv")
X1.2=read.csv(file="exampleM1Y20.1002AGE.csv")
M1=matrix(0,ncol=5,nrow=100)

M1=data.frame(M1)

for(i in 1:50)
{
  M1[i,]=X1[c(3,5,8,11,14),(i+1)]
  M1[(50+i),]=X1.2[c(3,5,8,11,14),(i+1)]
}
colnames(M1)=c("Age","Constant","log U","log V","log E")

write.csv(M1,file="M1Y20AGE.csv")
```

```
Y140DA=read.csv("M1Y20AGE.csv")

#
#
#

X2=read.csv(file="exampleM2Y20.5002AGE.csv")
X2.2=read.csv(file="exampleM2Y20.1002AGE.csv")
M2=matrix(0,ncol=5,nrow=100)

M2=data.frame(M2)

for(i in 1:50)
{
  M2[i,]=X2[c(3,5,8,11,14),(i+1)]
  M2[(50+i),]=X2.2[c(3,5,8,11,14),(i+1)]
}
colnames(M2)=c("Age","Constant","log U","log V","log E")

write.csv(M2,file="M2Y20AGE.csv")

Y240DA=read.csv("M2Y20AGE.csv")
#
#
#
setwd("C:/Users/KOCI 2/Desktop/Y60AGE_SEPT_5_19")
X3=read.csv(file="exampleM3Y60.5002AGE.csv")
X3.2=read.csv(file="exampleM3Y60.1002AGE.csv")
M3=matrix(0,ncol=5,nrow=100)

M3=data.frame(M3)

for(i in 1:50)
{
  M3[i,]=X3[c(3,5,8,11,14),(i+1)]
  M3[(50+i),]=X3.2[c(3,5,8,11,14),(i+1)]
}
colnames(M3)=c("Age","Constant","log U","log V","log E")

write.csv(M3,file="M3Y60AGE.csv")
```

```
Y340DA=read.csv("M3Y20AGE.csv")

#
#
#

X4=read.csv(file="exampleM4Y20.5002AGE.csv")
X4.2=read.csv(file="exampleM4Y20.1002AGE.csv")
M4=matrix(0,ncol=5,nrow=100)

M4=data.frame(M4)

for(i in 1:50)
{
  M4[i,]=X4[c(3,5,8,11,14),(i+1)]
  M4[(50+i),]=X4.2[c(3,5,8,11,14),(i+1)]
}
colnames(M4)=c("Age","Constant","log U","log V","log E")

write.csv(M4,file="M4Y20AGE.csv")

Y440DA=read.csv("M4DAY40AGE.csv")

#
#
#

X1=read.csv(file="exampleM5Y20.5002AGE.csv")
X1.2=read.csv(file="exampleM5Y20.1002AGE.csv")
M1=matrix(0,ncol=5,nrow=100)

M1=data.frame(M1)

for(i in 1:50)
{
  M1[i,]=X1[c(3,5,8,11,14),(i+1)]
  M1[(50+i),]=X1.2[c(3,5,8,11,14),(i+1)]
}
colnames(M1)=c("Age","Constant","log U","log V","log E")

write.csv(M1,file="M5Y20AGE.csv")
```

```
Y540DA=read.csv("M5DAY40AGE.csv")

#
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X6=read.csv(file="exampleM6DAY40.50AGE.csv")
X6.2=read.csv(file="exampleM6DAY40.100AGE.csv")
M6=matrix(0,ncol=5,nrow=100)

M6=data.frame(M6)

for(i in 1:50)
{
  M6[i,]=X6[c(3,5,8,11,14),(i+1)]
  M6[(50+i),]=X6.2[c(3,5,8,11,14),(i+1)]
}
colnames(M6)=c("Age","Constant","log U","log V","log E")

write.csv(M6,file="M6DAY40AGE.csv")

Y640DA=read.csv("M6DAY40AGE.csv")

#
#
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X7=read.csv(file="exampleM7DAY40.50AGE.csv")
X7.2=read.csv(file="exampleM7DAY40.100AGE.csv")
M7=matrix(0,ncol=5,nrow=100)

M7=data.frame(M7)

for(i in 1:50)
{
  M7[i,]=X7[c(3,5,8,11,14),(i+1)]
  M7[(50+i),]=X7.2[c(3,5,8,11,14),(i+1)]
}
colnames(M7)=c("Age","Constant","log U","log V","log E")

write.csv(M7,file="M7DAY40AGE.csv")
```

```
Y740DA=read.csv("M7DAY40AGE.csv")

setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X8=read.csv(file="exampleM8DAY40.50AGE.csv")
X8.2=read.csv(file="exampleM8DAY40.100AGE.csv")
M8=matrix(0,ncol=5,nrow=100)

M8=data.frame(M8)

for(i in 1:50)
{
  M8[i,]=X8[c(3,5,8,11,14),(i+1)]
  M8[(50+i),]=X8.2[c(3,5,8,11,14),(i+1)]
}
colnames(M8)=c("Age","Constant","log U","log V","log E")

write.csv(M8,file="M8DAY40AGE.csv")

Y840DA=read.csv("M8DAY40AGE.csv")

#
#
#
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X9=read.csv(file="exampleM9DAY20.50AGE.csv")
X9.2=read.csv(file="exampleM9DAY20.100AGE.csv")
M9=matrix(0,ncol=5,nrow=100)

M9=data.frame(M9)

for(i in 1:50)
{
  M9[i,]=X9[c(3,5,8,11,14),(i+1)]
  M9[(50+i),]=X9.2[c(3,5,8,11,14),(i+1)]
}
colnames(M9)=c("Age","Constant","log U","log V","log E")

write.csv(M9,file="M9DAY40AGE.csv")
```

```
Y940DA=read.csv("M9DAY40AGE.csv")
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_JULY_17_19")

X10=read.csv(file="exampleM10MDAY40.50AGE.csv")
X10.2=read.csv(file="exampleM10MDAY40.100AGE.csv")
M10=matrix(0,ncol=5,nrow=100)

M10=data.frame(M10)

for(i in 1:50)
{
  M10[i,]=X10[c(3,5,8,11,14),(i+1)]
  M10[(50+i),]=X10.2[c(3,5,8,11,14),(i+1)]
}
colnames(M10)=c("Age","Constant","log U","log V","log E")

write.csv(M10,file="M10DAY40AGE.csv")

Y1040DA=read.csv("M10DAY40AGE.csv")
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_JULY_17_19")

X11=read.csv(file="exampleM11DAy40.50AGE.csv")
X11.2=read.csv(file="exampleM11DAY40.100AGE.csv")
M11=matrix(0,ncol=5,nrow=100)

M11=data.frame(M11)

for(i in 1:50)
{
  M11[i,]=X11[c(3,5,8,11,14),(i+1)]
  M11[(50+i),]=X11.2[c(3,5,8,11,14),(i+1)]
}
colnames(M11)=c("Age","Constant","log U","log V","log E")

write.csv(M11,file="M11DAY40AGE.csv")
```

```
Y1140DA=read.csv("M11DAY40AGE.csv")
#
#

setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X12=read.csv(file="exampleM12DAY40.50AGE.csv")
X12.2=read.csv(file="exampleM12DAY40.100AGE.csv")
M12=matrix(0,ncol=5,nrow=100)

M12=data.frame(M12)

for(i in 1:50)
{
  M12[i,]=X12[c(3,5,8,11,14),(i+1)]
  M12[(50+i),]=X12.2[c(3,5,8,11,14),(i+1)]
}
colnames(M12)=c("Age","Constant","log U","log V","log E")

write.csv(M12,file="M12DAY40AGE.csv")

Y1240DA=read.csv("M12DAY40AGE.csv")
#
#
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X13=read.csv(file="exampleM13DAY40.50AGE.csv")
X13.2=read.csv(file="exampleM13DAY40.100AGE.csv")
M13=matrix(0,ncol=5,nrow=100)

M13=data.frame(M13)

for(i in 1:50)
{
  M13[i,]=X13[c(3,5,8,11,14),(i+1)]
  M13[(50+i),]=X13.2[c(3,5,8,11,14),(i+1)]
}
colnames(M13)=c("Age","Constant","log U","log V","log E")

write.csv(M13,file="M13DAY40AGE.csv")
```

```
Y1340DA=read.csv("M13DAY40AGE.csv")
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X14=read.csv(file="exampleM14DAY40.50AGE.csv")
X14.2=read.csv(file="exampleM14DAY40.100AGE.csv")
M14=matrix(0,ncol=5,nrow=100)

M14=data.frame(M14)

for(i in 1:50)
{
  M14[i,]=X14[c(3,5,8,11,14),(i+1)]
  M14[(50+i),]=X14.2[c(3,5,8,11,14),(i+1)]
}
colnames(M14)=c("Age","Constant","log U","log V","log E")

write.csv(M14,file="M14DAY40AGE.csv")

Y1440DA=read.csv("M14DAY40AGE.csv")
#
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X15=read.csv(file="exampleM15DAY40.50AGE.csv")
X15.2=read.csv(file="exampleM15MDAY40.100AGE.csv")
M15=matrix(0,ncol=5,nrow=100)

M15=data.frame(M15)

for(i in 1:50)
{
  M15[i,]=X15[c(3,5,8,11,14),(i+1)]
  M15[(50+i),]=X15.2[c(3,5,8,11,14),(i+1)]
}
colnames(M15)=c("Age","Constant","log U","log V","log E")

write.csv(M15,file="M15DAY40AGE.csv")

Y1540DA=read.csv("M15DAY40AGE.csv")
#
#
#
```

```
#
setwd("C:/Users/OLUSEGUNISMAIL/Desktop/Combined_files_MDA_MNDA_
JULY_17_19")

X16=read.csv(file="exampleM16DAY40.50AGE.csv")
X16.2=read.csv(file="exampleM16DAY40.100AGE.csv")
M16=matrix(0,ncol=5,nrow=100)

M16=data.frame(M16)

for(i in 1:50)
{
  M16[i,]=X16[c(3,5,8,11,14),(i+1)]
  M16[(50+i),]=X16.2[c(3,5,8,11,14),(i+1)]
}
colnames(M16)=c("Age","Constant","log U","log V","log E")

write.csv(M16,file="M16DAY40AGE.csv")

Y1640DA=read.csv("M16DAY40AGE.csv")
#
#
#

truth=c(-0.08,6,log(3),log(2),log(1))

CM=colnames(M2)

for(i in 1:5)
{

boxplot(Y040DA[,i],Y140DA[,i],Y240DA[,i],Y340DA[,i],Y440DA[,i],Y540DA[,i],Y640DA[,i],
Y740DA[,i],Y840DA[,i],Y940DA[,i],Y1040DA[,i],Y1140DA[,i],Y1240DA[,i],Y1340DA[,i],Y14
40DA[,i],Y1540DA[,i],Y1640DA[,i],main=CM[i])


  abline(h=truth[i],col=2)
}
```

Appendix B

Weighted Analysis of the JSLC2007

```
*M0-Unweighted
gen M0=1
*M1-Dwelling weight
gen M1=1/pr_dwell_ed
*M2-Enumeration district weight
gen M2=1/pr_ed_const
*M4-sampling weight from dwelling to ED to constituency
gen M4= (1/pr_dwell_ed)*(1/pr_ed_const)
logistic tmissing age
predict pi,
bysort psu_1:egen avgpi=mean(pi)
gen M6 =1/avgpi
table M6
encode educ1, generate(educ2)
encode es1, generate(es2)
encode ocg1, generate(ocg2)

xtmixed logtmeal|| parish:|| psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal|| parish:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal|| psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal|| parish:|| psu_1:,pweight(M4)
estat ic
estat icc


xtmixed logtmeal|| psu_1:,pweight(M4)
estat ic
estat icc
```

```
xtmixed logtmeal|| parish:|| psu_1:,pweight(M6)
estat ic
estat icc

xtmixed logtmeal age|| parish:|| psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal age|| parish:|| psu_1:,pweight(M4)
estat ic
estat icc

xtmixed logtmeal age|| parish:|| psu_1:,pweight(M6)
estat ic
estat icc


xtmixed logtmeal hhsize1|| parish:|| psu_1:,pweight(M0)
estat ic
estat icc


xtmixed logtmeal hhsize1|| parish:|| psu_1:,pweight(M4)
estat ic
estat icc

xtmixed logtmeal hhsize1|| parish:|| psu_1:,pweight(M6)
estat ic
estat icc

xtmixed logtmeal sex_female || parish:|| psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal sex_female || parish:|| psu_1:,pweight(M4)
estat ic
estat icc

xtmixed logtmeal sex_female|| parish:|| psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal sex_female || parish:|| psu_1:,pweight(M6)
estat ic
estat icc
```

```
xtmixed logtmeal i.educ2 || parish:|| psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal i.educ2 || parish:|| psu_1:,pweight(M6)
estat ic
estat icc


xtmixed logtmeal educ1_cxcgce educ1_degreeother || parish:||
psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal educ1_cxcgce educ1_degreeother || parish:||
psu_1:,pweight(M4)
estat ic
estat icc

xtmixed logtmeal educ1_cxcgce educ1_degreeother || parish:||
psu_1:,pweight(M6)
estat ic
estat icc


xtmixed logtmeal es1_out  es1_unemployed|| parish:|| psu_1:,pweight(M4)
estat ic
estat icc

xtmixed logtmeal es1_out  es1_unemployed || parish:|| psu_1:,pweight(M6)
estat ic
estat icc

xtmixed logtmeal ocg1_clerk ocg1_crtw ocg1_eo ocg1_lsom ocg1_nc
ocg1_pmoa ocg1_professionals ocg1_swsm ocg1_tap || parish:||
psu_1:,pweight(M0)
estat ic
estat icc


xtmixed logtmeal ocg1_clerk ocg1_crtw ocg1_eo ocg1_lsom ocg1_nc
ocg1_pmoa ocg1_professionals ocg1_swsm ocg1_tap || parish:||
psu_1:,pweight(M4)
estat ic
estat icc
```

```
xtmixed logtmeal ocg1_clerk ocg1_crtw ocg1_eo ocg1_lsom ocg1_nc
ocg1_pmoa ocg1_professionals ocg1_swsm ocg1_tap || parish:||
psu_1:,pweight(M6)
estat ic
estat icc

xtmixed logtmeal age hhsize1 educ1_cxcgce educ1_degreeother ocg1_clerk
ocg1_crtw ocg1_eo ocg1_lsom  ocg1_pmoa ocg1_professionals ocg1_swsm
ocg1_tap || parish:|| psu_1:,pweight(M6)
estat ic
estat icc

xtmixed logtmeal age hhsize1 educ1_cxcgce educ1_degreeother ocg1_clerk
ocg1_crtw ocg1_eo ocg1_lsom  ocg1_pmoa ocg1_professionals ocg1_swsm
ocg1_tap || parish:|| psu_1:,pweight(M4)
estat ic
estat icc

xtmixed logtmeal age hhsize1 i.sex2 i.educ2 i.es2 i.ocg2  || parish:||
psu_1:,pweight(M0)
estat ic
estat icc

xtmixed logtmeal age hhsize1 i.sex2 i.educ2 i.es2 i.ocg2  || parish:||
psu_1:,pweight(M6)
estat ic
estat icc
```

Appendix C

CHNS 1989

```
**MODELIV-D1-D2-D3-D4-D5-D6 & **MODELIV-E1-E2-E3-E4-E5-E6
**indcome has missing income
**indinc without missing
destring indcome, replace
gen loginA=ln(indcome)
destring indinc, replace
gen loginB=ln(indinc)

gen male = (gender==1)
gen female= (gender==2)
table male
table female

gen Nevermarried =(a8==1)
gen Married =(a8==2)
gen Divorced =(a8==3)
gen Widowed =(a8==4)
gen Separated =(a8==5)
gen Uknown =(a8==9)
gen Notmarried= Nevermarried + Divorced + Widowed + Separated + Uknown
table Married
table Notmarried

****Table 5.7
***MODEL IV-D1
xtmixed loginB|| t1:||t3:||t4:
estat ic

***MODEL IV-D2
xtmixed loginB age || t1:||t3:||t4:
estat ic


***MODEL IV-D3
xtmixed loginB age male || t1:||t3:||t4:
estat ic

***MODEL IV-D4
xtmixed loginB age male a11 || t1:||t3:||t4:
estat ic
```

```
***MODEL IV-D5
xtmixed loginB age male a11 Married || t1:||t3:||t4:
estat ic

***MODEL IV-D6
xtmixed loginB age a11  || t1:||t3:||t4:
estat ic

****Weighted Multilevel Model

logistic miss age a11 Married male
logistic miss age a11 Married female
logistic miss female
predict pi,

bysort t1:egen avgpi=mean(pi)
gen W_jkli =1/avgpi

***Table 5.8
***MODEL IV-E1
xtmixed loginA|| t1:||t3:||t4:,pweight(W_jkli)
estat ic

***MODEL IV-E2
xtmixed loginA age || t1:|| t3:|| t4:,pweight(W_jkli)
estat ic

***MODEL IV-E3
xtmixed loginA age male|| t1:|| t3:||t4:,pweight(W_jkli)
estat ic


***MODEL IV-E4
xtmixed loginA age male a11 || t1:|| t3:||t4:,pweight(W_jkli)
estat ic


***MODEL IV-E5
xtmixed loginA age male a11 Married || t1:|| t3:|| t4:,pweight(W_jkli)
estat ic

***MODEL IV-E6
xtmixed loginA age a11  || t1:|| t3:|| t4:,pweight(W_jkli)
estat ic
```

Appendix D

CHNS 2011

```
***MODEL III-F2
xtmixed loginB age || t1:||t4:
estat ic
***MODEL III-F3
xtmixed loginB age male || t1:||t4:
estat ic
***MODEL III-F4
xtmixed loginB age male a11 || t1:||t4:
estat ic
***MODEL III-F5
xtmixed loginB age male a11 Married || t1:||t4:
estat ic
***MODEL III-F6
xtmixed loginB male a11  || t1:||t4:
estat ic
****Weighted Multilevel Model
logistic miss age a11 Married male
logistic miss age a11 Married female
logistic miss female
predict pi,

bysort t1:egen avgpi=mean(pi)
gen W_jkli =1/avgpi
```

```
***Table 5.10
***MODEL III-G1
xtmixed loginA|| t1:||t4:,pweight(W_jkli)
estat ic

***MODEL III-G2
xtmixed loginA age || t1:|| t4:,pweight(W_jkli)
estat ic

***MODEL III-G3
xtmixed loginA age male|| t1:|| t3:||t4:,pweight(W_jkli)
estat ic


***MODEL III-G4
xtmixed loginA age male a11 || t1:||t4:,pweight(W_jkli)
estat ic
***MODEL III-G5
xtmixed loginA age male a11 Married || t1:|| t4:,pweight(W_jkli)
estat ic

***MODEL III-G6
xtmixed loginA male a11  || t1:|| t4:,pweight(W_jkli)
estat ic
```

Appendix E

R-CODE FOR DATA GENERATION SIMULATION

```
setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVTODAY")
```

20 PERCENT PROPORTION OF MISSING IN THE DEPENDENT VARIABLE

MISSINGNESS DEPENDS ON AGE

DATA GENERATION BY AGE OF RESPONDENTS

```
X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age


N=length(ser)
Npsu=max(psu)
Npar=max(parish)


# Parameters

M=100

# Creates missingness based on age

Mdelta0=1.9
Mdelta1=-0.03


loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))

mean(loMiss)

# mean(loMiss) gives the mean amount of missingness.

Current setup missingness decreases  with age.

sigU=3
sigV=2
sigE=1
```

```
beta0=6

beta1=-0.08

#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)
{

Upar=rnorm(Npar,0,sigU)
Vpsu=rnorm(Npsu,0,sigV)
E=rnorm(N,0,sigE)

U=Upar[parish]
V=Vpsu[psu]


Y[,i]=beta0+beta1*age+U+V+E
Mi[,i]=rbinom(N,1,loMiss)
# Mi records which observations are observed/missing
}
#


YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")


YD=data.frame(W)

colnames(YD)=c("serial","psu_1","parish","age",Ynam)

write.csv(YD,"sim_y76011age.csv")
```

20 PERCENT PROPORTION OF MISSING IN THE DEPENDENT VARIABLE

MISSINGNESS NOT DEPENDING ON AGE

```
X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age


N=length(ser)
Npsu=max(psu)
Npar=max(parish)


# Parameters

M=100

Miss=0.2

sigU=3
sigV=2
sigE=1


beta0=6

beta1=-0.08

#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)
{

Upar=rnorm(Npar,0,sigU)
Vpsu=rnorm(Npsu,0,sigV)
E=rnorm(N,0,sigE)

U=Upar[parish]
V=Vpsu[psu]
Y[,i]=beta0+beta1*age+U+V+E

Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi records which observations are observed/missing
}
#


YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")


YD=data.frame(W)

colnames(YD)=c("serial","psu_1","parish","age",Ynam)

write.csv(YD,"sim_y76011age.csv")
```

MISSINGNESS NOT DEPENDING ON AGE

DATA GENERATED BY SEX OF RESPONDENTS

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVTODAY/Simulation")

X=read.csv("JAM1994TODAY.csv")

File_Number=3

filename=paste("sim_y",File_Number,".csv",sep="")

Name="sex" # covariate dependent

ser=X$serial
psu=X$psu_1
parish=X$parish
variname=as.numeric(X$sex)-1 # covariate dependent

N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters

M=100 # Total number of simulations

Miss=0.2

sigU=3
sigV=2
sigE=1

beta0=6
beta1=5
#

Y=matrix(0,ncol=M,nrow=N)

for(i in 1:M)
{

Upar=rnorm(Npar,0,sigU)
Vpsu=rnorm(Npsu,0,sigV)
E=rnorm(N,0,sigE)

U=Upar[parish]
V=Vpsu[psu]
Y[,i]=beta0+beta1*variname+U+V+E
}
#

Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi

YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=variname
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish",Name,Ynam)
write.csv(YD,filename)
```

MISSINGNESS NOT DEPENDING ON AGE

DATA GENERATED BY AGE AND SEX OF RESPONDENTS

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVTODAY")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=45000

filename=paste("sim_y",File_Number,".csv",sep="")

Name="sex" # covariate dependent
ser=X$serial
psu=X$psu_1
parish=X$parish
variname=as.numeric(X$sex)-1 # covariate dependent
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)
# Parameters

M=100 # Total number of simulations

Miss=0.2

sigU=3
sigV=2
sigE=1

beta0=6
beta1=5
betta2=3
#

Y=matrix(0,ncol=M,nrow=N)

for(i in 1:M)
{

Upar=rnorm(Npar,0,sigU)
Vpsu=rnorm(Npsu,0,sigV)
E=rnorm(N,0,sigE)

U=Upar[parish]
V=Vpsu[psu]


Y[,i]=beta0+beta1*variname+betta2*age+U+V+E
}
#
```

```
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi

YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+5),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=variname
W[,5]=age
W[,6:(M+5)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","variname","age","Ynam")
write.csv(YD,filename)
```

Appendix F

R-CODE TO INVESTIGATE THE EFFECT INTRACLASS CORRELATION

BASED ON DATA SIMULATION -MISSING DEPENDS ON AGE

setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

```
X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age

N=length(ser)
Npsu=max(psu)
Npar=max(parish)


# Parameters
M=100
# Creates missingness based on age
Mdelta0=1.9
Mdelta1=-0.03
loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))
mean(loMiss)
# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.
sigU=3
sigV=2
sigE=1


beta0=6
beta1=-0.08


#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)
for(i in 1:M)

{

Upar=rnorm(Npar,0,sigU)
Vpsu=rnorm(Npsu,0,sigV)
E=rnorm(N,0,sigE)



U=Upar[parish]
V=Vpsu[psu]
Y[,i]=beta0+beta1*age+U+V+E
Mi[,i]=rbinom(N,1,loMiss)
# Mi records which observations are observed/missing
}
#
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM


Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD,"sim_y2021101911age.csv")
```

```
setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")


X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age

N=length(ser)
Npsu=max(psu)
Npar=max(parish)


# Parameters
M=100
# Creates missingness based on age


Mdelta0=1.9
Mdelta1=-0.03


loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))
mean(loMiss)


# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.

sigU=1
sigV=2
sigE=3


beta0=6
beta1=-0.08


#
Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)
for(i in 1:M)
{

  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]


  Y[,i]=beta0+beta1*age+U+V+E
  Mi[,i]=rbinom(N,1,loMiss)
  # Mi records which observations are observed/missing
}
#


YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD,"sim_y2021101912age.csv")
```

```
setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")


X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age

N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters

M=100

# Creates missingness based on age

Mdelta0=1.9
Mdelta1=-0.03
loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))

mean(loMiss)

# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.

sigU=5
sigV=4
sigE=3


beta0=6
beta1=-0.08


#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)

{

  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)


  U=Upar[parish]
  V=Vpsu[psu]


  Y[,i]=beta0+beta1*age+U+V+E

  Mi[,i]=rbinom(N,1,loMiss)

  # Mi records which observations are observed/missing

}

#

YM=Y

YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM


Ynam=0

for(i in 1:M) Ynam[i]=paste("y",i,sep="")



YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD,"sim_y2021101913age.csv")
```

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVTODAY")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=45034

filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=3
sigV=2
sigE=1
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)

  Vpsu=rnorm(Npsu,0,sigV)

  E=rnorm(N,0,sigE)

  U=Upar[parish]

  V=Vpsu[psu]

  Y[,i]=beta0+beta1*age+U+V+E

}

#

Mi=matrix(rbinom(N*M,1,Miss),ncol=M)

#Mi

YM=Y

YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)

W[,1]=ser

W[,2]=psu

W[,3]=parish

W[,4]=age

W[,5:(M+4)]=YM

Ynam=0

for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)

colnames(YD)=c("serial","psu_1","parish","age",Ynam)

write.csv(YD, filename)


#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")


X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")


File_Number=102020211
```

```
filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent



ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=3
sigV=2
sigE=1
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)




#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVTODAY")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=102020212
```

```
filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent


ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=1
sigV=2
sigE=3
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)
```

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=10202021
filename=paste("sim_y",File_Number,".csv",sep="")
Name="age" # covariate dependent
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=3
sigV=2
sigE=1
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)
```

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVTODAY")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=45034

filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=3
sigV=2
sigE=1
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
```

```
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)
```

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=102020211

filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent


ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=3
sigV=2
sigE=1
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
```

```
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)
```

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=102020212

filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent


ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=1
sigV=2
sigE=3
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
```

```
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)


#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")


X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
```

```
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age



N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters

M=100

# Creates missingness based on age

Mdelta0=0
Mdelta1=-0.03

loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))
mean(loMiss)

# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.

sigU=3
sigV=2
sigE=1

beta0=6
beta1=-0.08
```

```
#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)
{

  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)

  U=Upar[parish]
  V=Vpsu[psu]



  Y[,i]=beta0+beta1*age+U+V+E
  Mi[,i]=rbinom(N,1,loMiss)
  # Mi records which observations are observed/missing
}
#



YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
```

```
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD,"sim_y102020213age.csv")

#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age

N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100
# Creates missingness based on age
Mdelta0=0
Mdelta1=-0.03
loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))
mean(loMiss)
```

```
# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.

sigU=1
sigV=2
sigE=3

beta0=6
beta1=-0.08
#
Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)
{

  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)

  U=Upar[parish]
  V=Vpsu[psu]

  Y[,i]=beta0+beta1*age+U+V+E
  Mi[,i]=rbinom(N,1,loMiss)
  # Mi records which observations are observed/missing
}
#
```

```
YM=Y

YM[Mi==1]=NA


W=matrix(0,ncol=(M+4),nrow=N)

W[,1]=ser

W[,2]=psu

W[,3]=parish

W[,4]=age

W[,5:(M+4)]=YM


Ynam=0

for(i in 1:M) Ynam[i]=paste("y",i,sep="")


YD=data.frame(W)

colnames(YD)=c("serial","psu_1","parish","age",Ynam)

write.csv(YD,"sim_y102020214age.csv")

#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=1020202117

filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent


ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
```

```
sigU=3
sigV=2
sigE=1
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)


#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")

File_Number=1020202128

filename=paste("sim_y",File_Number,".csv",sep="")

Name="age" # covariate dependent


ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age
N=length(ser)
```

```
Npsu=max(psu)
Npar=max(parish)

# Parameters
M=100 # Total number of simulations
Miss=0.202412
sigU=1
sigV=2
sigE=3
beta0=6
beta1=-0.08

#
Y=matrix(0,ncol=M,nrow=N)
for(i in 1:M)
{
  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)
  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
}
#
Mi=matrix(rbinom(N*M,1,Miss),ncol=M)
#Mi
YM=Y
YM[Mi==1]=NA
W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM
Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")
YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD, filename)
```

```
#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age


N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters

M=100

# Creates missingness based on age

Mdelta0=0
Mdelta1=-0.03

loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))
mean(loMiss)

# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.

sigU=3
sigV=2
sigE=1

beta0=6
beta1=-0.08

#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)
{

  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)

  U=Upar[parish]
  V=Vpsu[psu]
  Y[,i]=beta0+beta1*age+U+V+E
  Mi[,i]=rbinom(N,1,loMiss)
  # Mi records which observations are observed/missing
}
#


YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD,"sim_y1020202139age.csv")


#setwd("C:/users/OLUSEGUNISMAIL/DESKTOP/CSVOCT192021")

X=read.csv("JAM1994_JULY_5_2019_WITHOUT_WEIGHT.csv")
ser=X$serial
psu=X$psu_1
parish=X$parish
age=X$age


N=length(ser)
Npsu=max(psu)
Npar=max(parish)

# Parameters

M=100

# Creates missingness based on age

Mdelta0=0
Mdelta1=-0.03

loMiss=exp(Mdelta0+Mdelta1*age)/(1+exp(Mdelta0+Mdelta1*age))
mean(loMiss)

# mean(loMiss) gives the mean amount of missingness. Current setup missingness decreases with age.
```

```
sigU=1
sigV=2
sigE=3

beta0=6
beta1=-0.08

#

Y=matrix(0,ncol=M,nrow=N)
Mi=matrix(0,nrow=N,ncol=M)

for(i in 1:M)
{

  Upar=rnorm(Npar,0,sigU)
  Vpsu=rnorm(Npsu,0,sigV)
  E=rnorm(N,0,sigE)

  U=Upar[parish]
  V=Vpsu[psu]


  Y[,i]=beta0+beta1*age+U+V+E
  Mi[,i]=rbinom(N,1,loMiss)
  # Mi records which observations are observed/missing
}
#


YM=Y
YM[Mi==1]=NA

W=matrix(0,ncol=(M+4),nrow=N)
W[,1]=ser
W[,2]=psu
W[,3]=parish
W[,4]=age
W[,5:(M+4)]=YM

Ynam=0
for(i in 1:M) Ynam[i]=paste("y",i,sep="")

YD=data.frame(W)
colnames(YD)=c("serial","psu_1","parish","age",Ynam)
write.csv(YD,"sim_y1020202149age.csv")
```

Appendix G

Table 5.16: Parameter estimates from unweighted multilevel models from the missing mechanism is MAR with missing rates of 20% , 40% and 60% for data simulated from age -only and sex-only covariate models

MISSING AT RANDOM (MAR)											
				20%			**40%**			**60%**	
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	β_1 Age	UP	-0.08000	-0.08016	0.20000	0.07217	-0.08018	0.22500	0.07221	-0.07997	0.03750	0.07201
		PS	-0.08000	-0.08010	0.12500	0.07212	-0.08015	0.18750	0.07219	-0.07999	0.01250	0.07201
		HS	-0.08000	-0.08027	0.33750	0.07228	-0.08020	0.25000	0.07224	-0.08010	0.12500	0.07216
		PHS	-0.08000	-0.08023	0.28125	0.07225	-0.07860	1.74538	0.07242	-0.08013	0.16375	0.07219
		IR	-0.08000	-0.08016	0.20000	0.07217	-0.08019	0.24250	0.07222	-0.07996	0.04500	0.07201
Y_{2jkl}	α_1 Sex	UP	5.00000	5.00522	0.10440	0.05698	4.98987	0.20260	0.07230	4.99610	0.07800	0.07970
		PS	5.00000	5.00653	0.13060	0.06266	4.98731	0.25380	0.08272	4.99800	0.04000	0.09846
		HS	5.00000	5.00652	0.13040	0.06609	4.99182	0.16360	0.08399	4.99480	0.10400	0.09295
		PHS	5.00000	5.00870	0.17400	0.07176	4.99069	0.18620	0.09282	4.99770	0.04600	0.11387
		IR	5.00000	5.00503	0.10060	0.05692	4.99020	0.19600	0.07206	4.99571	0.08580	0.08009

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.17: Parameter estimates from unweighted multilevel models from the missing mechanism is MNAR with missing rates of 20% , 40% and 60% for data simulated from age -only and sex-only covariate models

MISSING NOT AT RANDOM (MNAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	β_1 Age	UP	-0.08000	-0.07992	0.10000	0.07194	-0.08014	0.17250	0.07217	-0.07989	0.13750	0.07194
		PS	-0.08000	-0.07990	0.12250	0.07192	-0.08000	0.00000	0.07203	-0.08004	0.04500	0.07210
		HS	-0.08000	-0.07993	0.08625	0.07195	-0.08013	0.16750	0.07217	-0.07975	0.31125	0.07180
		PHS	-0.08000	-0.07990	0.12250	0.07193	-0.08002	0.03000	0.07206	-0.07994	0.07875	0.07200
		IR	-0.08000	-0.07990	0.12125	0.07194	-0.08014	0.17125	0.07216	-0.07989	0.14250	0.07194
Y_{2jkl}	α_1 Sex	UP	5.00000	4.99988	0.00240	0.09223	4.99467	0.10660	0.06040	5.01574	0.31480	0.09223
		PS	5.00000	4.99877	0.02460	0.09154	4.99460	0.10800	0.06821	5.01571	0.31420	0.09154
		HS	5.00000	4.99672	0.06560	0.10060	4.99104	0.17920	0.06954	5.02195	0.43900	0.10060
		PHS	5.00000	4.99559	0.08820	0.10267	4.99074	0.18520	0.06926	5.01902	0.38040	0.10267
		IR	5.00000	4.99968	0.00640	0.09240	4.99469	0.10620	0.06002	5.01580	0.31600	0.09240

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.18 Parameter estimates from weighted multilevel models(scaled method B) from the missing mechanism is MAR with missing rates of 20% , 40% and 60% for each data simulated from sex- age and multivariable covariate models

SCLAED METHOD B - MISSING AT RANDOM (MAR)												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	β_1	UP	-0.08000	-0.08016	0.20000	0.07217	-0.08018	0.22500	0.07221	-0.07997	0.03750	0.07201
		PHSB	-0.08000	-0.08022	0.27875	0.07225	-0.08025	0.30875	0.072281	-0.08010	0.12750	0.072013
		IRB	-0.08000	-0.08016	0.20500	0.07218	-0.08020	0.24625	0.0722257	-0.07996	0.05125	0.072003
Y_{2jkl}	α_1	UP	5.00000	5.00522	0.10440	0.05698	4.98987	0.20260	0.07230	4.99605	0.07900	0.07970
		PHSB	5.00000	5.00587	0.11740	0.06978	5.00067	0.01340	0.08015	4.99288	0.14240	0.10404
		IRB	5.00000	5.00507	0.10140	0.05678	4.99013	0.19740	0.07227	4.99580	0.08400	0.08009
Y_{3jkl}	λ_1	UP	5.00000	4.99870	0.02600	0.06111	4.98966	0.20680	0.05964	4.99225	0.15500	0.08342
		PHSB	5.00000	4.99392	0.12160	0.06772	4.98847	0.23060	0.07016	4.99220	0.15600	0.09757
		IRB	5.00000	4.99871	0.02580	0.06134	4.98999	0.20020	0.0597214	4.99229	0.15420	0.08338
	λ_2	UP	-0.08000	-0.07982	0.22000	0.00153	-0.08019	0.23750	0.00155	-0.079327	0.84125	0.00250
		PHSB	-0.08000	-0.07996	0.04500	0.00180	-0.08029	0.35625	0.00193	-0.079831	0.21125	0.00282
		IRB	-0.08000	-0.07983	0.21625	0.00154	-0.08020	0.24875	0.00155	-0.07934	0.82500	0.00250
Y_{4jkl}	τ_1	UP	-0.00780	-0.00804	3.10769	0.00193	-0.00791	1.39615	0.00231	-0.00765	1.93282	0.00285
		PHSB	-0.00780	-0.00795	1.94744	0.00230	-0.00793	1.70513	0.00266	-0.00757	2.89436	0.00342
		IRB	-0.00780	-0.00793	1.60769	0.00200	-0.00773	0.86795	0.00229	-0.00765	1.88423	0.00304
	τ_3	UP	-0.02500	-0.03161	26.44281	0.04828	-0.02472	1.12492	0.06602	-0.01943	22.27556	0.08408
		PHSB	-0.02500	-0.03123	24.90088	0.05913	-0.01762	29.50520	0.07778	-0.01761	29.55320	0.10627
		IRB	-0.02500	-0.02768	10.70384	0.04671	-0.02724	8.94680	0.06560	-0.01977	20.92624	0.10420

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PHSB=ED and Household selection weight using scaled Method B; IRB=Item non-response weight using scaled Method B

Table 5.19: Parameter estimates from weighted multilevel models(scaled method B) from the missing mechanism is MNAR with missing rates of 20% , 40% and 60% for each data simulated from sex- age and multivariable covariate models

SCALED METHOD B - MISSING NOT AT RANDOM (MNAR)												
				20%			40%			60%		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	β_1	UP	-0.08000	-0.08016	0.20000	0.07217	-0.08018	0.22500	0.07221	-0.07997	0.03750	0.07201
		PHSB	-0.08000	-0.07987	-0.16625	0.07189	-0.07998	0.02875	0.07201	-0.07996	0.05000	0.07202
		IRB	-0.08000	-0.07993	-0.08625	0.07195	-0.08014	0.17500	0.07217	-0.07988	0.14750	0.07193
Y_{2jkl}	α_1	UP	5.00000	4.99988	-0.00240	0.05911	4.99467	0.10660	0.06040	5.01574	0.31480	0.09223
		PHSB	5.00000	4.99646	-0.07080	0.06487	4.99529	0.09420	0.06921	5.01968	0.39360	0.10140
		IRB	5.00000	4.99963	-0.00740	0.05929	4.99468	0.10640	0.06009	5.01570	0.31400	0.09251
Y_{3jkl}	λ_1	UP	5.00000	5.01085	0.21700	0.05398	5.00473	0.09460	0.06083	4.99372	0.12560	0.07657
		PHSB	5.00000	5.00451	0.09020	0.06265	5.01025	0.20500	0.08079	4.99231	0.15380	0.08511
		IRB	5.00000	5.01092	0.21840	0.05396	5.00461	0.09220	0.06090	4.99338	0.13240	0.07669
	λ_2	UP	-0.08000	-0.08034	0.42750	0.00166	-0.07994	0.07000	0.00166	-0.08004	0.04625	0.00240
		PHSB	-0.08000	-0.08040	0.50000	0.00190	-0.07999	0.01875	0.00204	-0.07992	0.09750	0.00295
		IRB	-0.08000	-0.08035	0.43375	0.00167	-0.07995	0.06250	0.00165	-0.08004	0.04625	0.00241
Y_{4jkl}	τ_1	UP	-0.00780	-0.00786	0.74359	0.00188	-0.00809	3.76282	0.00213	-0.00771	1.19436	0.00286
		PHSB	-0.00780	-0.00786	0.71538	0.00214	-0.00801	2.71154	0.00252	-0.00886	13.62779	0.00319
		IRB	-0.00780	-0.00757	2.96538	0.00188	-0.00826	5.92821	0.00217	-0.00867	11.15859	0.00296
	τ_3	UP	-0.02500	-0.02113	15.46564	0.05411	-0.03152	26.09620	0.06543	-0.01896	24.17556	0.09031
		PHSB	-0.02500	-0.01646	34.15704	0.06706	-0.03063	22.53920	0.07907	-0.03183	27.32624	0.09942
		IRB	-0.02500	-0.02504	0.16800	0.05164	-0.03386	35.42000	0.06405	-0.03068	22.70888	0.08115

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PHSB=ED and Household selection weight using scaled Method B; IRB=Item non-response weight using scaled Method B

Table 5.20: Random Component – (U) estimates from weighted multilevel models from the missing mechanism is MAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex- age and multivariable covariate models

MISSING AT RANDOM (MAR)-U												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	**U**	UP	3.00000	2.79059	6.98043	0.62431	2.77502	7.49919	0.59267	2.75840	8.05339	0.62480
		PS	3.00000	2.88365	3.87817	0.57616	2.86711	4.42962	0.55151	2.85078	4.97410	0.57373
		HS	3.00000	2.79055	6.98156	0.62440	2.77547	7.48446	0.59268	2.75823	8.05896	0.62445
		PHS	3.00000	2.91063	2.97888	0.55428	2.89988	3.33724	0.55428	2.88340	3.88671	0.56829
		IR	3.00000	2.80679	6.44019	0.58146	2.79295	6.90175	0.58146	2.77519	7.49357	0.61664
Y_{2jkl}	U	UP	3.00000	2.89286	3.57140	0.68486	2.83333	5.55564	0.54833	2.82946	5.68467	0.61433
		PS	3.00000	2.99894	0.03520	0.67219	2.93701	2.09954	0.52285	2.93051	2.31638	0.56889
		HS	3.00000	2.89299	3.56688	0.68518	2.83286	5.57136	0.54866	2.82965	5.67835	0.61401
		PHS	3.00000	3.02042	0.68055	0.52591	2.95298	1.56727	0.52591	2.95579	1.47382	0.57892
		IR	3.00000	2.91044	2.98538	0.53976	2.84940	5.02000	0.53976	2.84647	5.11753	0.60459
Y_{3jkl}	U	UP	3.00000	2.91979	2.67352	0.67522	2.63838	12.05387	0.69843	2.87170	4.27656	0.62435
		PS	3.00000	3.01213	0.40429	0.64580	2.73394	8.86865	0.63793	2.96728	1.09068	0.59223
		HS	3.00000	2.91974	2.67537	0.67496	2.63749	12.08371	0.69853	2.87271	4.24291	0.62474
		PHS	3.00000	3.04023	1.34097	0.66521	2.74075	8.64163	0.62979	2.98073	0.64249	0.58957
		IR	3.00000	2.93634	2.12210	0.66879	2.65697	11.43443	0.68521	2.98970	0.34350	0.58702
Y_{4jkl}	U	UP	3.00000	2.86345	4.55169	0.67293	2.72059	9.31377	0.57988	2.71592	9.46935	0.58426
		PS	3.00000	2.95535	1.48844	0.62623	2.80926	6.35794	0.53140	2.81592	6.13588	0.53344
		HS	3.00000	2.86314	4.56198	0.67298	2.72051	9.31649	0.57970	2.75435	8.18834	0.57386
		PHS	3.00000	2.98337	0.55439	0.64646	2.83840	5.38662	0.52276	2.83979	5.34023	0.52276
		IR	3.00000	2.88044	3.98529	0.66439	2.73859	8.71356	0.56851	2.73552	8.81594	0.56851

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.21: Random Component – (U)estimates from weighted multilevel models from the missing mechanism is MNAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex- age and multivariable covariate models

MISSING AT RANDOM (MNAR)-U												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	**U**	UP	3.00000	2.79364	6.87881	0.61629	2.77502	7.49919	0.59267	2.75840	8.05339	0.62480
		PS	3.00000	2.88882	3.70585	0.57631	2.86711	4.42962	0.57631	2.85078	4.97410	0.57373
		HS	3.00000	2.79362	6.87940	0.61636	2.77547	7.48446	0.61636	2.75823	8.05896	0.62445
		PHS	3.00000	2.90971	3.00970	0.56892	2.90971	3.00970	0.56892	2.88340	3.88671	0.56829
		IR	3.00000	2.81065	6.31162	0.60665	2.79295	6.90175	0.60665	2.77519	7.49357	0.61664
Y_{2jkl}	U	UP	3.00000	2.74367	8.54445	0.61951	2.91563	2.81226	0.56826	2.82772	5.74274	0.64407
		PS	3.00000	2.83938	5.35402	0.56669	3.01499	0.49967	0.54326	2.92767	2.41085	0.60367
		HS	3.00000	2.74339	8.55373	0.61953	2.91528	2.82415	0.56816	2.82757	5.74754	0.64389
		PHS	3.00000	2.85233	4.92232	0.55221	3.00803	0.26772	0.55514	2.95245	1.58483	0.59217
		IR	3.00000	2.75983	8.00566	0.60982	2.93357	2.21430	0.56106	2.84491	5.16975	0.63533
Y_{3jkl}	U	UP	3.00000	2.75169	8.27693	0.62002	2.81763	6.07915	0.62977	2.86691	4.43635	0.62929
		PS	3.00000	2.85435	4.85486	2.91242	2.91242	2.91925	0.58849	2.96299	1.23361	0.59932
		HS	3.00000	2.75175	8.27494	2.81737	2.81737	6.08780	0.63001	2.86618	4.46058	0.62990
		PHS	3.00000	2.87759	4.08024	2.94546	2.94546	1.81795	0.58354	2.97320	0.89349	0.60358
		IR	3.00000	2.76906	7.69814	2.83583	2.83583	5.47246	0.62063	2.88610	3.79653	0.62126
Y_{4jkl}	U	UP	3.00000	2.72233	9.25580	0.57939	2.88236	3.92133	0.52686	2.68868	10.37739	0.60357
		PS	3.00000	2.81028	6.32394	0.53045	2.96001	1.33288	0.50942	3.02418	0.80596	0.64577
		HS	3.00000	2.72218	9.26060	0.57938	2.92060	2.64651	0.52087	2.91029	2.99038	0.63840
		PHS	3.00000	2.83986	5.33798	0.52232	3.01210	0.40318	0.50176	3.02238	0.74585	0.63224
		IR	3.00000	2.73998	8.66729	0.56815	2.89797	3.40098	0.52107	2.88461	3.84632	0.63677

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.22: Random Component – (V)estimates from weighted multilevel models from the missing mechanism is MAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex- age and multivariable covariate models

MISSING AT RANDOM (MAR)-V												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	**V**	UP	2.00000	2.00857	0.42850	0.10514	1.98272	0.86377	0.12605	1.97657	1.17128	0.12260
		PS	2.00000	1.89841	5.07936	0.15227	1.88268	5.86606	0.17510	1.87206	6.39685	0.18152
		HS	2.00000	2.03399	1.69963	0.10871	2.02013	1.00631	0.12396	2.03909	1.95468	0.12255
		PHS	2.00000	1.87424	6.28776	0.17484	1.86405	6.79764	0.19266	1.85858	7.07098	0.19735
		IR	2.00000	1.99196	0.40187	0.10455	1.96644	1.67798	0.12779	1.95963	2.01840	0.12629
Y_{2jkl}	V	UP	2.00000	1.99666	0.16699	0.11751	2.00100	0.05004	0.11747	1.98896	0.55202	0.12971
		PS	2.00000	1.89629	5.18554	0.15779	1.89786	5.10675	0.16646	1.87630	6.18491	0.17620
		HS	2.00000	2.02293	1.14635	0.11818	2.03906	1.95286	0.12223	2.05005	2.50248	0.13621
		PHS	2.00000	1.87958	6.02123	0.17530	1.88542	5.72896	0.18355	1.85910	7.04523	0.19698
		IR	2.00000	1.97935	1.03275	0.11825	1.98350	0.82477	0.11755	1.97180	1.41001	0.13212
Y_{3jkl}	V	UP	2.00000	1.99770	0.11524	0.10063	1.99792	0.10387	0.12271	2.02233	1.11633	0.11250
		PS	2.00000	1.88470	5.76513	0.15587	1.93038	3.48115	0.17136	1.91467	4.26627	0.14411
		HS	2.00000	2.02419	1.20958	0.10187	2.03491	1.74545	0.12562	2.08334	4.16713	0.13766
		PHS	2.00000	1.86629	6.68553	0.17228	1.90298	4.85076	0.17355	1.89319	5.34038	0.16578
		IR	2.00000	1.98157	0.92170	0.10204	1.98168	0.91598	0.12297	2.00505	0.25264	0.10988
Y_{4jkl}	V	UP	2.00000	1.99531	0.23431	0.11620	2.01754	0.87711	0.11024	2.01054	0.52703	0.12009
		PS	2.00000	1.88854	5.57323	0.16214	1.91407	4.29626	0.13747	1.89035	5.48268	0.15736
		HS	2.00000	2.02195	1.09772	0.11705	2.05584	2.79176	0.12079	2.06568	3.28403	0.13208
		PHS	2.00000	1.87292	6.35387	0.17729	1.90079	4.96067	0.15193	1.87705	6.14769	0.17235
		IR	2.00000	1.97865	1.06764	0.11717	2.00085	0.04253	0.10876	1.99383	0.30869	0.11883

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.23: Random Component – (V)estimates from weighted multilevel models from the missing mechanism is MNAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex- age and multivariable covariate models

MISSING AT RANDOM (MNAR)-V												
				20%			**40%**			**60%**		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	RMSE
Y_{1jkl}	**V**	UP	2.00000	1.98517	0.74146	0.10889	2.00133	0.06672	0.12633	1.98619	0.69053	0.11403
		PS	2.00000	1.87706	6.14697	0.16665	1.87706	6.14697	0.16647	1.86933	6.53363	0.17331
		HS	2.00000	2.01200	0.59991	0.10723	2.01200	0.59991	0.13048	2.04740	2.36998	0.11958
		PHS	2.00000	1.85444	7.27796	0.18956	1.87194	6.40288	0.18869	1.85204	7.39782	0.18961
		IR	2.00000	1.96925	1.53749	0.11166	1.98436	0.78200	0.12645	1.96898	1.55124	0.11720
Y_{2jkl}	V	UP	2.00000	2.01278	0.63882	0.10356	2.00100	0.05004	0.11747	1.99843	0.07874	0.13539
		PS	2.00000	1.90061	4.96939	0.15031	1.89786	5.10675	0.16646	1.89001	5.49968	0.17971
		HS	2.00000	2.03878	1.93878	0.10866	2.03906	1.95286	0.12223	2.05998	2.99904	0.14399
		PHS	2.00000	1.88667	5.66657	0.16178	1.88542	5.72896	0.18355	1.87429	6.28549	0.19840
		IR	2.00000	1.99592	0.20376	0.10213	1.98350	0.82477	0.11755	1.98079	0.96038	0.13574
Y_{3jkl}	V	UP	2.00000	1.99754	0.12319	0.11391	2.00413	0.20669	0.12513	2.00740	0.36979	0.11445
		PS	2.00000	1.89437	5.28148	0.16045	1.89831	5.08460	0.16416	1.89471	5.26467	0.15799
		HS	2.00000	2.02380	1.18995	0.11511	2.04115	2.05733	0.12975	2.06903	3.45128	0.13090
		PHS	2.00000	1.87735	6.13229	0.17898	1.88006	5.99705	0.17994	1.87842	6.07920	0.17222
		IR	2.00000	1.98168	0.91608	0.11377	1.98676	0.66200	0.12447	1.98974	0.51289	0.11422
Y_{4jkl}	V	UP	2.00000	1.99531	0.23431	0.11620	1.96439	1.78028	0.13535	1.99968	0.01579	0.12277
		PS	2.00000	1.88854	5.57323	0.16214	1.85322	7.33914	0.19630	1.88627	5.68654	0.16404
		HS	2.00000	2.02195	1.09772	0.11705	2.00364	0.18213	0.13213	2.05378	2.68918	0.11879
		PHS	2.00000	1.87292	6.35387	0.17729	1.82746	8.62706	0.22269	1.87161	6.41957	0.17150
		IR	2.00000	1.97865	1.06764	0.11717	1.94753	2.62359	0.13932	1.97502	1.24911	0.11632

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight

Table 5.24: Random Component – (E) estimates from weighted multilevel models from the missing mechanism is MNAR with missing rates of 20% , 40% and 60% for each data simulated from age-only; sex-only; sex- age and multivariable covariate models

MISSING AT RANDOM (MNAR)-E												
				20%			40%			60%		
DGM	**PI**	**WM**	**TV**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**	**AE**	**PB**	**RMSE**
Y_{1jkl}	**E**	UP	1.00000	0.99941	0.05857	0.02033	0.99823	0.17727	0.02175	0.99730	0.27025	0.02885
		PS	1.00000	1.00008	0.00777	0.02163	0.99769	0.23113	0.02614	0.99623	0.37688	0.03249
		HS	1.00000	0.94513	5.48745	0.05861	0.92616	7.38421	0.07675	0.88555	11.44501	0.11865
		PHS	1.00000	0.99965	0.03498	0.02311	1.00026	0.02567	0.02770	0.99344	0.65648	0.03785
		IR	1.00000	0.99942	0.05840	0.02045	0.99832	0.16808	0.02181	0.99726	0.27393	0.02893
Y_{2jkl}	E	UP	1.00000	0.99771	0.22861	0.01928	1.00121	0.12051	0.02401	0.99948	0.05192	0.02905
		PS	1.00000	0.99777	0.22334	0.02083	1.00237	0.23671	0.02597	0.99873	0.12691	0.03158
		HS	1.00000	0.94512	5.48835	0.05852	0.92731	7.26903	0.07696	0.88745	11.25451	0.11642
		PHS	1.00000	0.99806	0.19441	0.02329	0.99913	0.08733	0.02354	0.99791	0.20925	0.03467
		IR	1.00000	0.99758	0.24180	0.01935	1.00149	0.14909	0.02412	0.99929	0.07129	0.02906
Y_{3jkl}	E	UP	1.00000	0.99771	0.22861	0.01928	1.00121	0.12051	0.02401	0.99948	0.05192	0.02905
		PS	1.00000	0.99777	0.22334	0.02083	1.00237	0.23671	0.02597	0.99873	0.12691	0.03158
		HS	1.00000	0.94512	5.48835	0.05852	0.92731	7.26903	0.07696	0.88745	11.25451	0.11642
		PHS	1.00000	0.99806	0.19441	0.02329	0.99913	0.08733	0.02354	0.99791	0.20925	0.03467
		IR	1.00000	0.99758	0.24180	0.01935	1.00149	0.14909	0.02412	0.99929	0.07129	0.02906
Y_{4jkl}	E	UP	1.00000	0.99548	0.45197	0.01847	0.99171	0.82910	0.02359	0.98645	1.35494	0.03403
		PS	1.00000	0.99389	0.61110	0.01906	0.99010	0.99041	0.02542	0.97871	2.12910	0.03531
		HS	1.00000	0.94112	5.88814	0.06193	0.91883	8.11740	0.08436	0.87113	12.88666	0.13098
		PHS	1.00000	0.99273	0.72710	0.02256	0.98893	1.10673	0.02881	0.97540	2.45967	0.03809
		IR	1.00000	0.99549	0.45127	0.01856	0.99177	0.82305	0.02369	0.98282	1.71769	0.02963

DGM = Data generation model; PI= Parameter of interest; WM =Weighting methods; TV = True value; AE=Average Estimate for parameter of interest; PB=Percent bias; RMSE=Root mean square error; UP= Unweighted parameter; PS=ED selection weight; HS =Household selection weight ; PHS=ED and Household selection weight; IR=Item non-response weight