INTERCENSAL UPDATING OF SMALL AREA ESTIMATES

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

at Massey University, Palmerston North, New Zealand

Marissa Cinco Isidro

2010

To Inay and Itay

Abstract

Small area estimation (SAE) involves fitting statistical models to generate statistics for areas where the sample size of the survey data is insufficient for generating precise estimates. A recent application of SAE techniques is in estimating local level poverty measures in Third World countries necessary for aid allocation and monitoring of the Millennium Development Goals (MDGs). The SAE technique commonly known as ELL method (Elbers et al., 2003) is extensively implemented by the World Bank in collaboration with national statistical agencies in most Third World countries. This technique generates estimates by fitting a linear mixed model to household level income or consumption using the survey and census data. The ELL method differs in various ways from the mainstream SAE techniques, two of which are emphasized in this thesis: (1) the ELL model does not include area level effects and (2) the model fitting technique follows a non-standard weighted generalized least squares (GLS).

Under the ELL method the survey and the census data are assumed to have been conducted at the same time period, hence generating updated estimates of poverty measures during non-census years is a problem. The method for SAE updating developed in this thesis is called the Extended Structure Preserving Estimation (ESPREE) method, an extension of the classical SAE technique called the structure preserving estimation (SPREE) method - an approach to SAE based on a categorical data analysis framework. The ESPREE method is structured within a generalized linear model (GLM) framework and uses information from the most recent survey and pseudo-census (census replicates) data to generate updated small area estimates under a superpopulation.

The World Bank in collaboration with the National Statistical Coordination Board in the Philippines has conducted an intercensal updating project using an ELL-based method requiring time invariant variables. Comparison of the estimates generated from the ELL-based and ESPREE updating method revealed substantial differences. The ESPREE method but not the ELL updating method generated unbiased estimates. An in-country validation exercise conducted in the Philippines supported the view that ESPREE based estimates, besides having theoretical advantages, also conformed better to local experts' opinion on current poverty levels.

Acknowledgements

Undertaking my PhD research requires the help of countless people and institutions, without them nothing could have been accomplished. I would like to acknowledge some of them, because to enumerate all of them is impossible.

I am very much privileged to have been accepted as a student of two of the prominent international figures and experts in small area estimation of poverty statistics in developing countries. I have learned so much from them, their valuable suggestions and constant guidance made this thesis possible. They have been very supportive in all aspects of my post graduate student life, from ensuring that my research was in the right direction, to helping me secure sufficient financial support. They are like my two most dependable sherpas in my quest to reach the highest academic mountain. To Steve Haslett and Geoff Jones, thank you so much. It has been an amazing journey!

I am so fortunate to belong to the Department of Statistics in Massey University, composed of high caliber statisticians who are very friendly and helpful, headed by the ever smiling and very approachable subject leader, Martin Hazelton. These statisticians have inspired and helped me in so many ways. To all of you, thank you very much.

Just like any research endeavor, I benefited so much from the assistance of various institutions - provision of the necessary data, academic research, and financial related support. I would like to extend my gratitude to the head and staff of the following institutions: the World Food Program (WFP), Bureau for Asia in Thailand; the National Statistics Office (NSO) and the National Statistical Coordination Board (NSCB) in the Philippines; Education New Zealand through the New Zealand Postgraduate Study Abroad Awards (NZPSAA); the Institute of Fundamental Sciences in Massey University; the International Center of Excellence for Education in Mathematics (ICE-EM) in Australia; the Cavite State University (CvSU) and the various provincial government offices in the Ilocos region of the Philippines.

In the ups and downs of my postgraduate student life, it was comforting to know that I was not alone in my struggles. My fellow Statistics postgraduate students from different parts of the world have been a source of encouragement and have made me feel " normal" most of the time! My utmost gratefulness goes to all of them especially to Ting Wang, my officemate in the writing up room.

Some of the data sets required for my research were outputs from a collaborative project of the World Bank with the NSCB. I am so fortunate to have good friends and former colleagues in NSCB whose untiring support for all my requests made my access to the needed data easier. To Dette Balamban, Glennie Amoranto, Tess Almarines, Joseph and Mildred Addawe, and Art Martinez *marami pong salamat sa inyong lahat.*

Map generation for the poverty statistics produced from my research requires some programming skills and knowledge of the necessary computer software. I managed to

save time generating the required poverty maps with the help of a former classmate and a very good friend. To Irvin Samalca, *daghang salamat Kyo!*

I would not have gone this far academically had it not been for my former lecturers and mentors who have inspired and inculcated in me an interest in the wonders of the field of Statistics. I am forever thankful to Jacqueline Guarte, Lisa Bersales, Ana Tabunda, and Erniel Barrios.

Doing research can sometimes be tiring, some rest and recreation are therefore necessary. I am so grateful to have found new friends who have helped kept my sanity and made my leisure time meaningful, full of fun and laughter. They have also opened up opportunities for me to experience different cultures, learn new language and most of all enjoy and feast on delicious food! I am very grateful to Jojo Roldan and family, Olive Pimentel and family, Cheryl Fernandez, Mimi Dogimab, Andree Wallace, May Nawanuparatsakul, Poy Theerasin, Emily Kawabata, Lala Komalawati, Leela Awaludin, Phine and Bong Flores, Edith and Roy Meeking.

Long time friends who have always been there through thick and thin. Friends who have helped me in various ways while I was doing my research. To Venus Bermudo, Badet Montana, Beng Umali, Adel Rivera, Miriam Du, Aris Magallanes, Marian Baclayon and Sanae Tacata, thank you very much.

Living *Down Under* has been bearable because I never felt alone. Up in the *Great White North* I have my best friend, who has constantly encouraged me in so many times that I was on the verge of giving up. To my husband, Phelan, saying *do tze* is not sufficient to express my gratefulness for everything you have done for me. I am also indebted to both our families for their untiring love, prayers and moral support in the past three years.

Above all, I would like to acknowledge the *Greatest Statistician* of all time. The only one who knows the true value of all the parameters on earth. Thank you so much for the gift of wisdom, good health and strength.

Palmerston North, New Zealand   Marissa Isidro
31 August 2010

Table of Contents

List of Tables

List of Figures

Introduction

Reliable and timely local level information on various concerns (e.g., local economic situation, business opportunities, health conditions, educational needs, and poverty incidence) or characteristics of the population are needed for planning, policymaking, and decision making both in the public and private sectors. The adoption of a new paradigm of governance, decentralization of state power, by most countries has contributed to the worldwide increase in demand for local level information. The new paradigm involves the transfer of state/national responsibilities or functions from central government to sub-national or local government units (e.g., provinces and municipalities), or from central agencies/offices to regional bodies or branch offices, or to non-governmental or private organizations (Larbi, 1999). Implementation of local plans and programs as well as evaluation and monitoring would need local level information.

This local level information may be available from national surveys that are usually conducted by national statistical agencies; however, the level of precision may not be acceptable to be used for planning and policy making purposes. This is because the majority of those surveys are designed to generate reasonably accurate "direct" or survey-based estimates for the characteristics or parameters of interest only up to the second administrative level (e.g., for the Philippines - national and regional). Other sub-national administrative levels in the Philippines include provinces (usually there are at least four provinces in each region) which are composed of municipalities. A municipality is composed of villages called barangays, these villages are generally the primary sampling units (PSUs) in the survey design for national surveys such as the Family Income and Expenditure Survey (FIES). Thus, while the total sample is widely distributed over the country, the sample within smaller geographic units or local administrative levels (e.g. municipalities in the Philippines) is usually very small or even nonexistent in some instances. With very small sample sizes at the local level, the direct estimates (if possible) generated usually have very large standard errors. If more precise estimates are desired to be computed directly from the survey for local levels, a large-scale survey should be conducted, which would mean an increase in the survey funds (which are usually limited) needed and could further lengthen the data processing time.

The term "local area" or "small area" is not restricted to local administrative level or sub-national government units mentioned above (which is the small area of particular importance in this research). Small area in general refers to subsets of the population; these subsets may refer to geographic subdivisions (e.g., county, states, and provinces); demographic subdivisions (e.g., race-sex-occupation) within a large geographical area; or other groupings for which the samples from (national) surveys are too small to provide estimates with acceptable levels of precision (Rao, 2003).

The demand for reliable local level or small area information has led to the development of a range of estimation techniques, commonly known as "small area estimation" methods. This set of statistical techniques generally allows the generation of more reliable small area estimates without adding much burden to the limited resources of most national or private statistical agencies. There are already a number of reviews written on developments in small area estimation, the most recent ones being the reviews written by Ghosh and Rao (1994), Rao (1999), the book written by Rao (2003), and Jiang and Lahiri (2006). One interesting development is the formulation of a framework for all small area models: the general linear regression model (Marker, 1999) and generalized linear model (Noble et al., 2002). A common framework facilitates comparison of small area models, as well as examination and understanding of model assumptions in the different methods.

Small area estimation techniques generate more reliable estimates at the local level by using "indirect" estimators (model-based estimators, as opposed to "direct" estimators which as mentioned are survey-based) that "borrow strength" by using values of the variable of interest, y, from related areas and/or time periods and thus increase the "effective" sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas and/or time periods through the supplementary information related to y, such as recent census counts and current administrative records (Rao, 2003).

There are various applications of small area estimation in the literature - agriculture, education, business, health, employment and socio-economics. One of the most recent applications is the estimation of poverty statistics at the local level (e.g. municipalities) in Third World countries. Poverty statistics have gained much importance with poverty alleviation being the first among the eight Millennium Development Goals (MDGs) embodied in the United Nations (UN) Millennium Declaration adopted by the heads of state around the world during the 2000 UN Millennium Summit (UN-website, 2009).

Poverty is a very complex multidimensional concern: there is no single definition and method of measurement available. In this research, we adhere to the meaning of poverty that is used by most economists, i.e., households are considered to be in poverty if their income falls below some income threshold called the poverty line. Chambers (2006) described this as income-poverty, and it is the definition adopted by the World Bank in the implementation of their poverty mapping projects carried out in collaboration with national statistical agencies and used, for example, for monitoring progress towards the MDGs. Sometimes expenditure-based poverty estimates are used instead to measure economic poverty, and in public health related contexts, measures such as standardized weight for age, height for age and weight for height (underweight, stunting and wasting, respectively) in children under five years of age are used, e.g. in Bangladesh (Haslett and Jones, 2004) and Nepal (Haslett and Jones, 2006).

Information on consumption and/or income that is used to determine poverty measures is generally obtained through national surveys (e.g. Family Income and Expenditure Survey (FIES), in the Philippines), with which sample households are asked to answer detailed questions on their spending habits and sources of income. Such surveys are conducted once every three years in most countries. The FIES which is conducted once every three years, is an example of surveys that only allow acceptable level of precision of estimates up to the second administrative level. Hence, survey-based poverty statistics in the Philippines (as in other Third World countries) have an acceptable level of precision at the regional level (second administrative level). However, for policy makers to properly target assistance and interventions to the neediest communities and households, more disaggregated poverty statistics are needed.

The need for more disaggregated poverty statistics in Third World countries useful for aid allocation and monitoring of poverty alleviation projects led to the development of a small area estimation methodology for poverty measures appropriate to the available data in Third World countries. In 2003, Elbers, Lanjouw and Lanjouw (ELL) proposed a method specifically designed for poverty statistics in Third World countries at lower geographical or administrative levels, commonly referred to as ELL methodology. This methodology has been adopted by the World Bank in their poverty mapping projects in most Third World countries (in collaboration with national statistical agencies). Some modifications have been made in the implementation of this methodology in other countries, e.g., computations of the household level variance (Fujii, 2003) and in the estimation of parameters and selection of small area models, see for example the implementation in Bangladesh (Haslett and Jones, 2004), Philippines (Haslett and Jones, 2005) and Vietnam (Minot et al., 2003), and the inclusion or consideration of small area level random effects (e.g., Haslett and Jones (2006)).

The ELL methodology combines the sample survey and census data to come up with small area poverty estimates. This is a very useful small area estimation technique in the World Bank's effort to generate poverty statistics necessary for aid allocation and poverty monitoring. Their methodology however has some issues related to its theoretical underpinning as will be illustrated in Chapter 2. In using the census and survey data for generating small area estimates, the ELL method assumes that the two data sets are gathered at the same time period. This assumption is particularly important since the variable of interest (income/consumption or poverty status) is not measured in the census; hence, the model for income is formulated using the survey data and is then applied to the census data. In most countries especially in Third World countries, a census is only conducted once in every ten years. This poses a problem in the generation of updated small area estimates during non-census years or intercensal years. It is therefore important to develop an intercensal updating method for small area estimates of poverty measures in Third World countries to provide policymakers and other stakeholders with an updated estimate of poverty measures, hence this research.

This thesis has two main parts: first, is the background on small area estimation methods and the discussion of the issues of the ELL method which is used for generating small area estimates of poverty measures in Third World countries as well as the ways in which the method could be improved; and second, is the method developed for generation of the updated small area estimates of poverty measures in Third World countries, i.e. generation of small area estimates of poverty during non-census or intercensal years or years when there is a new survey data available but there is no new census and its application using the Philippine data. The first part is presented in Chapters 1 and 2 while the second part is presented in Chapters 3 to 8. Specific description of the different Chapters are presented below.

Chapter 1 provides an overview of small area estimation methods. Since there are already various reviews written on small area estimation and the various models and methods used in different applications, the main focus of this Chapter is on models and estimation methods relevant to poverty estimation in Third World countries. The ELL method mentioned above is basically using a linear mixed model for income/expenditure, hence estimation procedures for the linear mixed model as applied to small area estimation are reviewed in detail.

Chapter 2 contains a detailed description of the ELL method, in which the model, parameter estimation method and generation of small area estimates are presented. As pointed out above, there are some issues with the ELL income or expenditure model and its parameter estimation method; these issues are discussed and alternative approaches are suggested which include the "standard" parameter estimation methods for linear mixed models presented in Chapter 1. The different parameter estimation methods are compared and are illustrated using data from a survey in the Philippines. The majority of the materials in this Chapter will be published in Statistics Canada publication *Survey Methodology*, Catalogue 12-001-XIE2010002, December 2010, vol. 36 no. 2.

Chapter 3 gives an overview of intercensal updating for small area estimates. Different methods and applications are described including methods used by demographers in updating census counts during non-census years. Updating methods for small area estimates of poverty measures due to Lanjouw and van der Wiede (2006), Jitsuchon and Lanjouw (2005) and Hoogeveen et al. (2003) are also presented. These updating methods for poverty measures are either extensions or modifications of the original ELL method. The discussion of these methods highlights their most recent application to intercensal updating project of the World Bank in collaboration with the national statistical agencies in selected Third World countries.

Chapter 4 presents a detailed discussion of the structure preserving estimation (SPREE) method. A thorough discussion of the background of the SPREE method is necessary since the intercensal updating method proposed in this research is based on an extension of the SPREE method. Recent modifications and extensions of the SPREE method are also presented highlighting the most recent method proposed by Zhang and Chambers (2004) aimed at reducing the bias due to the assumption of a fixed census data under SPREE. Zhang and Chambers (2004) method involves the use of Generalized Linear Structural Models (GLSMs) which require the estimation of a parameter called the proportionality coefficient that accounts for changes in the census data, leading to a reduction in bias.

Chapter 5 is devoted to the details of the proposed intercensal updating method called extended SPREE (ESPREE). This method is basically an extension of the SPREE method formulated to generate updated small area estimates, in the sense that it allows for variability in the census data by using a superpopulation, as opposed to the classical SPREE method wherein the census is assumed fixed. The ESPREE method therefore extends SPREE by allowing for variation both in the survey and the census projections. Moreover, the ESPREE method generates updated small area estimates by fitting a GLM model to each set of census data drawn from the superpopulation and then adjusting those parameters that can be accurately estimated from the survey. The classical SPREE method on the other hand, uses the iterative proportional fitting (IPF) algorithm to generate small area estimates. The ESPREE method is also compared with the ELL-based updating methods and the GLSMs.

Chapter 6 addresses the problem of estimating the variances or standard errors of the updated small area estimates based on the ESPREE method. The variance estimation methods discussed here are general methods which includes linearization and replication methods. These methods could also be used in any estimation problem that uses data from a complex survey and a census. Under ESPREE, the census data is assumed stochastic (i.e., pseudo-census data) hence there are two sources of variation for the ESPREE based updated estimates - the variability from the survey and the pseudo-census data.

Chapter 7 provides an application of the ESPREE method using the Philippine data. The ESPREE method is used to generate updated estimates of poverty incidence using the 2003 survey data and the 2000 census data. The updated small area poverty incidence estimates are compared with the estimates generated by the ELL-based updating method of Lanjouw and van der Wiede (2006). The ESPREE updated estimates appear to be unbiased compared with the ELL-based updated estimates: at the provincial and regional levels, the ESPREE method generated updated small area estimates that have smaller coefficient of variation due to more explicit modelling and closer to direct survey-based estimates than the ELL-based estimates.

Chapter 8 presents the results of a validation study conducted in one of the regions in the Philippines. Differences between updated small area estimates generated using the ESPREE method and the ELL-based updating method were observed and as pointed out above, the ELL-based updated estimates seemed to be biased. With funding assistance from New Zealand Postgraduate Study Abroad Awards (NZPSAA), the Ilocos region in the Philippines has been visited to conduct a validation exercise. In general the estimates generated from the ESPREE method performed better on the ground than the ELL-based updating method, i.e., the key informants perception tend to agree more with the ESPREE-based than the ELL-based updating estimates.

Chapter 9 contains a summary of the research findings and some concluding remarks as well as recommendations on further research for small area estimation for poverty measures in Third World countries as well as on the generation of updated small area estimates of poverty measures.

Chapter 1

Small Area Estimation

1.1 Introduction

Small area estimation (SAE) methods can be considered to belong to two subdivisions: techniques with *implicit* models and those with *explicit* models. For those techniques using implicit models, the underlying models are known however the estimates are not generated by specifying an appropriate model explicitly. On the other hand, for techniques that employ explicit models, the underlying models are specified explicitly in order to generate the required estimates. Classical SAE methods belong to the group of SAE techniques that use implicit models such as *traditional demographic* methods where demographic variables are used to generate population or census updates during non-census years, and the *indirect domain* estimation methods, for example the "synthetic method" which uses survey-based or direct estimates to generate more precise small area estimates. These classical methods, as will be discussed in more detail in Chapter 3 (the Chapter devoted to intercensal updating) do not account for between area variation. On the other hand, the more recent SAE methods that use explicit models account for between area variation and are classified into *area level* models - area level auxiliary variables are used for the formulation of the small area model, and *unit level* models - unit level auxiliary variables are available.

Regardless of the SAE method used, either using implicit or explicit models and area level or unit level models for small area estimation, the models used can be put into the framework of generalized linear mixed models (GLMMs). Hence, in this Chapter an overview of the GLMMs is given (Section 1.2). A common framework for small area estimation models, aimed at facilitating comparison and selection of optimal method for a particular small area estimation problem, was initiated by Marker (1999) and followed by Noble et al. (2002). Marker (1999) proposed using linear mixed models, while Noble et al. (2002) consider the generalized linear models (GLMs). However, there are other small area models that cannot be considered under the class of GLMs but are covered in the broader set of models - the GLMMs.

A comprehensive discussion of various small area models and estimation methods is presented by Rao (2003) and in the review articles of Ghosh and Rao (1994) and Rao (1999). Since the field of small area estimation is so broad and this thesis focuses on small area estimation of poverty measures in developing or Third World countries, only SAE models related to the most widely implemented SAE method for poverty measures in Third World countries, the ELL (Elbers et al., 2003) method, are reviewed in detail. Discussion of the ELL method is presented in the next Chapter. The implementation of the ELL method basically involves fitting a unit level linear mixed model (a special case of GLMM) to income or consumption data, hence a review of the different estimation methods available for unit level models with a linear mixed model structure is necessary. Description of the linear mixed model is presented in Section 1.3 while the different parameter estimation methods are discussed in Section 1.4. Estimation techniques include Estimated Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB) and Hierarchical Bayes (HB) methods. Modifications to the basic methods used to generate design consistent estimates are presented in Section 1.5 and a summary for the Chapter is given in Section 1.6.

1.2 Framework for Small Area Models

Generalized linear models (GLMs) are a class of models introduced by Nelder and Wedderburn (1972) which represents a group of fixed effects regression models for various response variables which may for example be continuous, binary or count variables. Hence, GLMs can be considered as a generalization of the classical linear models. A particular GLM has three components, namely, the *random component*, *systematic component* and the *link function*. A GLM relates the distribution of the response variable $\mathbf{Y}$ (random component) to the predictor variables $\mathbf{X}$ (systematic component) through the link function. The GLM is therefore defined by the choices of the random component, systematic component and the link function.

For the GLM, the distribution of the response variable or random component is generated from the exponential family of distributions, which are those probability distributions, parameterized by θ and ϕ, that have density functions or probability mass functions (depending on whether the distribution is continuous or discrete) which can be expressed in the form:

$$f_{\mathbf{Y}}(y;\theta,\phi) = exp(\frac{\tilde{a}(y)\tilde{b}(\theta)+\tilde{c}(\theta)}{\jmath(\phi)} + \tilde{d}(y,\phi)) \tag{1.1}$$

The parameter θ is related to the mean of the distribution while the parameter ϕ, called the dispersion parameter, is related to the variance of the distribution. The form of the functions $\tilde{a}$, $\tilde{b}$, $\tilde{c}$ and $\tilde{d}$ are assumed known. The exponential family includes the normal, binomial and Poisson distributions, among others. In small area estimation for poverty measures in Third World countries only the normal distribution has been extensively used so far, with variables such as income or expenditure at household level being transformed to normality (usually by a log function). Although the parameters of interest are nonlinear functions of income or expenditure such as poverty incidence, gap and severity which are described in the next Chapter.

The systematic component of the GLM specifies that the predictor variables $\boldsymbol{X}$ relate to the level of the response or dependent variable $\boldsymbol{Y}$ as a linear combination of the predictor variable, $\boldsymbol{\eta} = \boldsymbol{X\beta}$. The link function (assumed to be a monotonic differentiable function) then relates $\boldsymbol{\eta}$ to the mean of $\boldsymbol{Y}$ (i.e., $\boldsymbol{\mu}$), via the function g so that $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. Hence, the form of the generalized linear model is:

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta} \tag{1.2}$$

One of the distributions of particular interest in modeling poverty incidence (or strictly speaking poverty prevalence) is the Poisson distribution (related discussion is presented in Chapter 4). This is used to model count data (number of poor and non-poor in a small area cross-classified by some relevant auxiliary variables). Under this distribution, the usual link function is the logarithm, i.e. $\boldsymbol{\eta} = log(\boldsymbol{\mu})$. The variance function is proportional to the mean, $Var(\mathbf{Y}) = \boldsymbol{\phi\mu}$ where the dispersion parameter $\boldsymbol{\phi}$ is generally a vector of ones. A case of Poisson with overdispersion or *quasipoisson* occurs when $\boldsymbol{\phi}$ is greater than one. We note however that under the ELL method, poverty incidence are generated via household level predictions of income or expenditure model (linear mixed model).

In many small area applications, the units on which observations or measurements of the variable of interest have been made are not necessarily independent of each other. For example, incomes of households that are clustered together or located in the same village tend to be more similar than those far apart. This type of data is commonly known as *clustered data* and a similar data structure exists for *longitudinal data*, *repeated measures* and *multi-level data*. The GLM model which only considers fixed effects and therefore assumes that all observations are independent of each other would not be sufficient to account for the correlations present in the data. For proper analysis of the data, a cluster or area effect which is assumed to be random is included in the model. The model then contains both fixed and random effects. The set of models containing both random and fixed effects is called *Generalized Linear Mixed Models* (GLMMs). Under this set of models, the equation for $\boldsymbol{\eta}$ then becomes:

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\upsilon}$$

where $\boldsymbol{\upsilon}$ is a vector of random effects and $\boldsymbol{Z}$ is similar to (and can be a subset of) $\boldsymbol{X}$, the model matrix. The random effects account for the correlations present in the data that is not captured by the auxiliary variables or covariates. When the link function is the identity function and the distribution of the error processes is normal, then the GLMM model is equivalent to a linear mixed model. The linear mixed model (LMM) is described in more detail in the next Section.

1.3 The Linear Mixed Model

In general LMMs have the following structure:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\upsilon} + \boldsymbol{e} \tag{1.3}$$

where $\boldsymbol{Y}$ is the $N \times 1$ vector of the response variable (or variable of interest), $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known $N \times \dot{p}$ and $N \times q$ design matrices of full rank, $\boldsymbol{\beta}$ is $\dot{p} \times 1$, $\boldsymbol{\upsilon}$ is $q \times 1$ and $\boldsymbol{e}$ is $N \times 1$. The design matrices are required to be of full rank in order to ensure that the parameters are not linearly dependent upon one another. In cases where categorical variables are used, the number of categories less one is considered for modelling to simplify algebra by satisfying the full rank assumption. The random effects $\boldsymbol{\upsilon}$ and the error component $\boldsymbol{e}$ are assumed to be independently distributed with means $\mathbf{0}$ and covariance matrices $\boldsymbol{G} = V(\boldsymbol{\upsilon})$ and $\boldsymbol{R} = V(\boldsymbol{e})$, respectively. The variance of $\boldsymbol{Y}$ which is a function of $\boldsymbol{G}$ and $\boldsymbol{R}$ is as follows $V(\boldsymbol{Y}) = \boldsymbol{ZGZ}' + \boldsymbol{R}$. In most small area applications, the first component of $V(\boldsymbol{Y})$ represents the between small areas portion of the covariance structure while the second component represents the within small area covariance.

A more specific linear mixed model that resembles the model used in the ELL method is the linear mixed model with a block-diagonal covariance matrix structure. Under this model, $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_A')'$ denotes the vector of the response variable, and the design or model matrix is denoted by $\boldsymbol{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_A')'$ and $\boldsymbol{Z} = \text{diag}\{\mathbf{Z}_a\}$, respectively. Here, $a = 1, ..., A$ where A is the number of independent sets of observation (which could be clusters or small areas in the context of small area estimation) in the population. The vector of random effects is $\boldsymbol{v} = (\boldsymbol{v}_1', \ldots, \boldsymbol{v}_A')'$ while $\boldsymbol{e} = (\boldsymbol{e}_1', \ldots, \boldsymbol{e}_A')'$ is the vector of random errors. Here, $\mathbf{Y}_a$ is a $N_a \times 1$ vector, $\mathbf{X}_a$ is $N_a \times \dot{p}$ and $\mathbf{Z}_a$ is an $N_a \times q_a$ matrix, while $\boldsymbol{v}_a$ is $q_a \times 1$ and $\boldsymbol{e}_a$ is an $N_a \times 1$ vector where $\sum N_a = N$ and $\sum q_a = q$; when q_a is constant say $q_a = q_1$ for all a, then $q = Aq_1$. In addition, $\boldsymbol{R} = \text{diag}\{\boldsymbol{R}_a\}$ and $\boldsymbol{G} = \text{diag}\{\boldsymbol{G}_a\}$ so that $V(\boldsymbol{Y}) = \boldsymbol{V} = \text{diag}\{\boldsymbol{V}_a\}$ has a block-diagonal structure, with $\boldsymbol{V}_a = \mathbf{Z}_a \boldsymbol{G}_a \mathbf{Z}_a' + \boldsymbol{R}_a$. Using the new notation, the linear mixed model with block-diagonal covariance matrix may be decomposed into A submodels as

$$\mathbf{Y}_a = \mathbf{X}_a \boldsymbol{\beta} + \mathbf{Z}_a \boldsymbol{v}_a + \boldsymbol{e}_a \tag{1.4}$$

The parameters of interest here are the linear combinations of the regression parameters $\boldsymbol{\beta}$ and the realization of $\boldsymbol{v}_a$ as follows:

$$\mu_a = \boldsymbol{l}_a' \boldsymbol{\beta} + \mathbf{c}_a' \boldsymbol{v}_a \tag{1.5}$$

for specified vectors of constants $\boldsymbol{l}_a$ and $\mathbf{c}_a$; in small area estimation μ_a could be the small area mean or total.

1.4 Small Area Estimation Techniques

There are basically three approaches to generate small area estimates through mixed models that are discussed in the literature e.g., 1) the classical prediction approach called the empirical best linear unbiased prediction (EBLUP), 2) empirical bayes (EB) and 3) the hierarchical bayes (HB). The discussion of the three methods in Sections 1.4.2 to 1.4.4 focuses on the "basic linear mixed model" because its structure is similar to the income/expenditure model used in the ELL method (as will be shown in the next Chapter). For simplicity of exposition, the data on the variable of interest is assumed to have been collected by conducting simple random sampling from each area. This survey design however is not necessarily the design used to collect the data for estimating poverty measures in most Third World countries. An example from the Philippines is presented in the next Chapter.

The "basic linear mixed model" mentioned above is as follows:

$$Y_{ah} = \mathbf{X}_{ah}\boldsymbol{\beta} + \upsilon_a + e_{ah} \tag{1.6}$$

Here, the variable of interest is Y_{ah} and is assumed to be related to the element-specific auxiliary data $\mathbf{X}_{ah} = (X_{ah1}, \ldots, X_{ah\dot{p}})$ and that $a = 1, \ldots, A$, $h = 1, \ldots, N_a$, $\boldsymbol{\beta} = (\beta_1, \ldots \beta_{\dot{p}})'$ is $\dot{p} \times 1$ vector of regression parameters and N_a is the number of population units or households in the ath small area. It is also assumed that the random effects υ_a are independent and identically distributed (`iid`) with expected value zero and variance σ_υ^2 and are independent of the unit errors $e_{ah} = k_{ah}\tilde{e}_{ah}$ with known constants k_{ah}, to allow for heteroscedasticity, and $\tilde{e}_{ah}$ are random errors assumed `iid` with mean zero and variance σ_e^2. Normality of the υ_a's and $\tilde{e}_{ah}$'s are also often assumed. Note that $\dot{p}$ is used here and is different from p that may be used to denote either probability or proportion in other Chapters. In matrix notation, the model (1.6) is as follows:

$$\boldsymbol{Y}_a = \mathbf{X}_a\boldsymbol{\beta} + \upsilon_a\mathbf{1}_{N_a} + \boldsymbol{e}_a \tag{1.7}$$

where $\mathbf{1}_{N_a}$ is an $N_a \times 1$ vector of ones. Model (1.7) is a special case of model (1.4). Under this linear mixed model, the parameters of interest are the small area means $\bar{Y}_a$ or the totals Y_a which are not necessarily the parameters of interest in the models formulated for the ELL method which may instead be focused on nonlinear functions of Y_a. As will be discussed in the next Chapter the parameters of interest are rather functions of the mean say $\theta_a = f(\bar{Y}_a)$, specifically, non-linear functions.

1.4.1 Framework for SAE method

Rao (2003) discussed an approach to estimate small area means or totals when the sampling rate ($\mathfrak{s}_a = n_a/N_a$) is not negligible, i.e. sample size (n_a) in small area a is large relative to the population size (N_a). In this situation a model-based estimate can be improved by considering both the model-based estimate and the survey-based estimate of say the small area mean. The small area mean can be expressed as follows:

$$\bar{Y}_a = \mathfrak{s}_a \bar{y}_a + (1 - \mathfrak{s}_a)\bar{y}_a^*$$

where $\bar{y}_a$ is the sample mean (mean of the sampled observations) and $\bar{y}_a^*$ is the mean of the non-sampled observations, y_{ah}^*, of the ath area. To generate the estimate of the small area mean, $\hat{\bar{Y}}_a$, under the population model (1.6), the unobserved values, y_{ah}^*, are replaced by the model-based estimator $(\mathbf{x}_{ah}^{*\prime}\hat{\boldsymbol{\beta}} + \hat{v}_a)$, where $\mathbf{x}_{ah}^*$ is the value of the auxiliary variables associated with the unobserved variable of interest y_{ah}^*, so that,

$$\hat{\bar{Y}}_a = \mathfrak{s}_a \bar{y}_a + (1 - \mathfrak{s}_a)(\bar{\mathbf{x}}_a^* \hat{\boldsymbol{\beta}} + \hat{v}_a) \tag{1.8}$$

where $\bar{\mathbf{x}}_a^*$ is the mean of $\mathbf{x}_{ah}^*$. If the sampling rate $\mathfrak{s}_a$ is negligible then the estimate of the small area means tend to depend more heavily on the model-based estimator, which is equivalent to having the small area means estimated by $\hat{\bar{Y}}_a = \bar{\boldsymbol{X}}_a \hat{\boldsymbol{\beta}} + \hat{v}_a$ where $\bar{\boldsymbol{X}}_a$ is the ath population mean of the set of auxiliary variables. Three methods of generating the model-based estimates of small area means are described in the next Sections.

We note that the small area estimator in equation (1.8) is related to the composite (shrinkage) estimator proposed by Longford (1999). Under the composite estimation method, a more precise small area estimate of the variable of interest is generated by combining the unstable (i.e., less precise) direct small area (e.g., sub-national or other local domains) estimates $\bar{y}_a$ and the more stable estimate say for example the national level direct estimate $\bar{y}$ as follows:

$$\hat{\bar{Y}}_a = \varphi_a \bar{y} + (1 - \varphi_a)\bar{y}_a \tag{1.9}$$

where the area or domain level coefficient φ_a is determined by minimizing the mean square error - MSE($\hat{\bar{Y}}_a; \bar{Y}_a$). The estimator shrinks toward the more stable estimate $\bar{y}$ when the direct small area estimates $\bar{y}_a$ are less stable, in the same manner that the small area estimator in (1.4.1) shrinks to the model-based estimator when the direct small area estimates are less precise, i.e, large N_a relative to n_a (which is generally small when available). The composite estimator in its 'most basic form' is simpler, as it does not involve fitting a model for the variable of interest as it only uses the direct small area estimate $\bar{y}_a$ and the more stable direct estimate for higher domain level $\bar{y}$. However, direct small area estimates are not always available, so modification to the basic shrinkage estimator involves the formulation of models to generate the small area estimates. Using models for the small area estimates ($\bar{y}_a$) leads to a framework similar to the approach by Rao (2003) described above. Composite estimation has also been extended to cover estimation for several variables, when it is called *multivariate shrinkage estimation.*

1.4.2 Empirical Best Linear Unbiased Prediction (EBLUP)

The development of the best linear unbiased prediction (BLUP) method for prediction of mixed effects dates back to the early works of C. R. Henderson in the late 1940s as outlined by Jiang and Lahiri (2006). As described by Robinson (1991), the BLUP estimates of the realized value of random variables υ are: (i)*linear* in the sense that they are linear functions of the data, y; (ii)*unbiased* because the estimate's expected value is equal to the average value of the quantity being estimated; and (iii) *best*, since among all the estimators that are both linear and unbiased (i.e., satisfies (i) and (iii)), the BLUP estimators have the minimum mean squared error.

The BLUP method in the context of the linear mixed model with block-diagonal covariance structure is presented here, based on the detailed description provided by Rao (2003). As pointed out in the previous Section, if the population sizes (N_a) of the small areas are sufficiently large, i.e., the sampling rates are negligible then we can take the ath small area mean as $\mu_a = \bar{\boldsymbol{X}}_a\boldsymbol{\beta} + \upsilon_a$. To generate the estimates of the small area means $\hat{\mu}_a$, we assume that the sample data y_{ah} and $\mathbf{x}_{ah}$ obey the population model in (1.6), so that we will have:

$$y_{ah} = \mathbf{x}_{ah}\boldsymbol{\beta} + \upsilon_a + e_{ah} \tag{1.10}$$

where $h = 1, \ldots, n_a$; $a = 1, \ldots, A$, or in matrix notation:

$$\boldsymbol{y}_a = \mathbf{x}_a\boldsymbol{\beta} + \upsilon_a\mathbf{1}_{n_a} + \boldsymbol{e}_a \tag{1.11}$$

We note the change in notation here from $\boldsymbol{Y}_a$ to $\boldsymbol{y}_a$. In general under small area estimation we are dealing with sample and population or census data for the variable of interest (Y) and auxiliary data (X), so for clarity we are using upper case roman letters (Y_{ah}, $\boldsymbol{Y}_a$, $\boldsymbol{Y}$) to refer to the population data of the variable of interest and lower case letter for the sample (y_{ah}, $\boldsymbol{y}_a$, $\boldsymbol{y}$). Similar notational differences hold for the auxiliary variables, (X_{ah}, $\boldsymbol{X}_a$, $\boldsymbol{X}$) for the population and ($\mathbf{x}_{ah}$, $\mathbf{x}_a$, $\mathbf{x}$) for the sample. We also note that μ_a and $\bar{Y}_a$ are used interchangeably to represent the small area mean parameter and their corresponding estimators are denoted by $\hat{\mu}_a$ and $\hat{\bar{Y}}_a$, respectively.

Given model (1.10) or (1.11), the BLUP estimator of the small area mean μ_a is as follows:

$$\hat{\mu}_a = \bar{\boldsymbol{X}}_a\hat{\boldsymbol{\beta}} + \gamma_a(\bar{y}_{ad} - \bar{\mathbf{x}}_{ad}\hat{\boldsymbol{\beta}}) \tag{1.12}$$

where $\bar{y}_{ad}$ and $\bar{\mathbf{x}}_{ad}$ are weighted means, $\bar{y}_{ad} = \sum_h d_{ah}y_{ah}/d_{a.}$ and $\bar{\mathbf{x}}_{ad} = \sum_h d_{ah}\mathbf{x}_{ah}/d_{a.}$ such that $d_{ah} = k_{ah}^{-2}$ and $d_{a.} = \sum_h d_{ah}$, k_{ah} is the constant that allows for heteroscedasticity in the unit level error term ($\tilde{e}_{ah}$), i.e. $e_{ah} = k_{ah}\tilde{e}_{ah}$ as pointed out in the description of the basic linear mixed model above, and $\gamma_a = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/d_{a.})$. Equation (1.12) is a special case of equation (1.5) where $\boldsymbol{l}_a' = \bar{\boldsymbol{X}}_a$ and $\boldsymbol{c}_a = 1$. Assuming full rank for $\boldsymbol{V}_a$, the best linear unbiased estimator (BLUE) of the fixed parameter $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\sum_a \mathbf{x}_a'\boldsymbol{V}_a^{-1}\mathbf{x}_a)^{-1}(\sum_a \mathbf{x}_a'\boldsymbol{V}_a^{-1}\boldsymbol{y}_a)$$

where

$$\boldsymbol{V}_a^{-1} = \frac{1}{\sigma_e^2}[\mathrm{diag}_h\{d_{ah}\} - \frac{\gamma_a}{d_{a.}}\boldsymbol{d}_a\boldsymbol{d}_a']$$

The BLUP estimator given in equation (1.12) depends on the variance components σ_e^2 and σ_v^2 which are usually unknown. By replacing the variance components with their estimates, an empirical BLUP (EBLUP) is obtained.

The author employed the estimators of variance components proposed by Henderson (1953), derived by computing the mean squares through a conventional least squares analysis of non-orthogonal data (e.g. method of fitting-of-constants), then equating the mean squares to their expectations to solve for the unknown variances. These variance component estimates are claimed to generate unbiased estimates even though some elements of the model are correlated. The Henderson's estimators of σ_e^2 and σ_v^2

are as follows:

$$\hat{\sigma}_e^2 = (n - A - \dot{p} + 1)^{-1} \sum_{a=1}^{A} \sum_{h=1}^{n_a} \hat{\varepsilon}_{ah}^2 \tag{1.13}$$

where $\{\hat{\varepsilon}_{ah}\}$ are residuals from the ordinary least squares (OLS) regression of $y_{ah} - \bar{y}_a$ on $\{x_{ah1} - \bar{x}_{a.1}, \ldots, x_{ah\dot{p}} - \bar{x}_{a.\dot{p}}\}$ and $(\bar{y}_a, \bar{x}_{a.1}, \ldots, \bar{x}_{a.\dot{p}})$ are the sample means in the ath group;

$$\hat{\sigma}_v^2 = n_*^{-1}[\sum_{a=1}^{A} \sum_{h=1}^{n_a} \hat{u}_{ah}^2 - (n - \dot{p})\hat{\sigma}_e^2] \tag{1.14}$$

where $n_* = n - tr[(\mathbf{x}'\mathbf{x})^{-1} \sum_{a=1}^{A} n_a^2 \bar{\mathbf{x}}_a \bar{\mathbf{x}}_a']$, $\{\hat{u}_{ah}\}$ are the residuals from the OLS regression of y_{ah} on $\{x_{ah1}, \ldots, x_{ah\dot{p}}\}$ and $n = \sum_{a=1}^{A} n_a$. Alternative estimators of the variance components may be considered, such as the maximum likelihood or the restricted maximum likelihood (REML) method by assuming normality.

1.4.3 Empirical Bayes

As described by Rao (2003), the Empirical Bayes (EB) method is implemented by first obtaining the posterior density, $f(\boldsymbol{\mu}|\boldsymbol{y}, \dot{\boldsymbol{\lambda}})$ (note that the dot here is used to signify the difference from the notation 'λ without a dot' used in Chapter 4 to represent variable effects) in a loglinear model, of the small area (random) parameters of interest, $\boldsymbol{\mu}$, given the data $\boldsymbol{y}$, using the conditional density, $f(\boldsymbol{y}|\boldsymbol{\mu}, \dot{\boldsymbol{\lambda}}_1)$, of $\boldsymbol{y}$ given $\boldsymbol{\mu}$ and the density of $\boldsymbol{\mu}$, $f(\boldsymbol{\mu}|\dot{\boldsymbol{\lambda}}_2)$ where $\dot{\boldsymbol{\lambda}} = (\dot{\boldsymbol{\lambda}}_1', \dot{\boldsymbol{\lambda}}_2')'$ denotes the vector of model parameters. Having obtained the posterior density, the model parameters, $\dot{\boldsymbol{\lambda}}$, are estimated from the marginal density, $f(\boldsymbol{y}|\dot{\boldsymbol{\lambda}})$, of $\boldsymbol{y}$. Then, the estimated posterior density, $f(\boldsymbol{\mu}|\boldsymbol{y}, \hat{\dot{\boldsymbol{\lambda}}})$, is used to make inferences about $\boldsymbol{\mu}$, where $\hat{\dot{\boldsymbol{\lambda}}}$ is an estimator of $\dot{\boldsymbol{\lambda}}$.

The EB method can be considered to combine the frequentist and Bayesian approaches to estimation, in the sense that the density of $\boldsymbol{\mu}$, as pointed out by Rao (2003), can be interpreted as a prior density on $\boldsymbol{\mu}$, but it is actually a part of the postulated model on $(\boldsymbol{y}, \boldsymbol{\mu})$ which can be validated from the data unlike the subjective priors on model parameters, $\dot{\boldsymbol{\lambda}}$, used in hierarchical Bayes (HB) approach. Details of the HB approach is described in the next Section.

The details of the EB procedure for the basic linear mixed model, i.e. a special case of the linear mixed model with block diagonal covariance structure (model 1.7), are presented here. As in the previous Section, we assume that the sample data $\boldsymbol{y}_a$ and $\mathbf{x}_a$ obey the population model (1.7) and assuming independence and normality, we will have the model, $\boldsymbol{y}_a|\upsilon_a \overset{\text{ind}}{\sim} N(\mathbf{x}_a\boldsymbol{\beta} + \upsilon_a\mathbf{1}_{n_a}, \boldsymbol{R}_a)$ such that $\upsilon_a \overset{\text{ind}}{\sim} N(0, \sigma_\upsilon^2)$ where $\boldsymbol{R}_a = \sigma_e^2\text{diag}\{k_{ah}^2\}$ and $a = 1, \ldots, A$ and $h = 1, \ldots, n_a$. Similarly, we are interested in estimating $\mu_a = \boldsymbol{l}'_a\boldsymbol{\beta} + \mathbf{c}'_a\boldsymbol{\upsilon}_a$ such that $\mathbf{c}_a = 1$, $\boldsymbol{\upsilon}_a = \upsilon_a$, $\boldsymbol{l}'_a = \bar{\boldsymbol{X}}_a$ and we note that $a = 1, \ldots, A$ for the parameter μ_a.

To generate the required parameter estimates, first, we will get the posterior density of μ_a given $\boldsymbol{y}_a$ which is as follows:

$$\mu_a|\boldsymbol{y}_a, \boldsymbol{\beta}, \sigma_\upsilon^2, \sigma_e^2 \overset{\text{ind}}{\sim} N(\hat{\mu}_a^B, g_{1a}) \tag{1.15}$$

where $g_{1a} = \gamma_a(\sigma_e^2/d_{a.})$ and $\hat{\mu}_a^B$ is the conditional expectation of μ_a given $\boldsymbol{y}_a$, $\boldsymbol{\beta}$ and $(\sigma_\upsilon^2, \sigma_e^2)$ i.e.,

$$\hat{\mu}_a^B = \hat{\mu}_a^B(\boldsymbol{\beta}, \sigma_\upsilon^2, \sigma_e^2) = E(\mu_a|\boldsymbol{y}_a, \boldsymbol{\beta}, \sigma_\upsilon^2, \sigma_e^2) = \boldsymbol{l}'_a\boldsymbol{\beta} + \hat{\upsilon}_a^B \tag{1.16}$$

which is considered as the Bayes predictor of μ_a and $\hat{\upsilon}_a^B = E(\upsilon_a|\boldsymbol{y}_a, \boldsymbol{\beta}, \sigma_\upsilon^2, \sigma_e^2) = \gamma_a(\boldsymbol{y}_a - \mathbf{x}_a\boldsymbol{\beta})$ such that $\boldsymbol{V}_a = \boldsymbol{R}_a + \sigma_\upsilon^2\mathbf{1}_{n_a}\mathbf{1}'_{n_a}$ and γ_a is as defined in equation (1.12)in the previous Section. Then, the model parameters $\boldsymbol{\beta}$ and $(\sigma_\upsilon^2, \sigma_e^2)$ in the Bayes predictor $\hat{\mu}_a^B$ are estimated from the marginal distribution

$$\boldsymbol{y}_a \overset{\text{ind}}{\sim} N(\mathbf{x}_a\boldsymbol{\beta}, \boldsymbol{V}_a) \tag{1.17}$$

using maximum likelihood (ML) or restricted maximum likelihood (REML) estimation methods, see Cressie (1992) for details. Having the estimates $\hat{\boldsymbol{\beta}}$ and $(\hat{\sigma}_\upsilon^2, \hat{\sigma}_e^2)$ the empirical Bayes predictor or EB estimator of the small area mean μ_a is as follows:

$$\hat{\mu}_a^{EB} = \boldsymbol{l}'_a\hat{\boldsymbol{\beta}} + \hat{\upsilon}_a^B \tag{1.18}$$

This estimator is identical to the EBLUP estimator in the previous Section with $\boldsymbol{l}'_a = \bar{\boldsymbol{X}}_a$.

1.4.4 Hierarchical Bayes

As mentioned in the previous Section, the hierarchical Bayes (HB) approach assumes a subjective prior distribution on the model parameters $\dot{\boldsymbol{\lambda}}$, that is $f(\dot{\boldsymbol{\lambda}})$ is specified.

The posterior density $f(\boldsymbol{\mu}|\boldsymbol{y})$ of the small area (random) parameters of interest $\boldsymbol{\mu}$, given the data $\boldsymbol{y}$, is obtained by using the conditional densities $f(\boldsymbol{y}|\boldsymbol{\mu}, \dot{\boldsymbol{\lambda}}_1)$ and $f(\boldsymbol{\mu}|\dot{\boldsymbol{\lambda}}_2)$ and combining with the subjective prior on the parameters $\dot{\boldsymbol{\lambda}} = (\dot{\boldsymbol{\lambda}}_1', \dot{\boldsymbol{\lambda}}_2')'$ using Bayes theorem. Inferences are made based on the posterior density: a particular parameter is estimated by its posterior mean and its posterior variance gives a measure of precision of the estimator. Evaluation of the posterior density and its associated parameters, e.g. the posterior mean and variance, involves multi-dimensional integrations, which can be very complicated. Computational difficulties can be overcome by using Markov chain Monte Carlo (MCMC) methods, for detailed discussion see Brooks (1998). MCMC methods provide an alternative to direct numerical integration such that the required posterior quantities (posterior mean and variance) are approximated from the samples generated from the posterior distribution.

One of the common sampling algorithms under the MCMC methods is the Gibbs sampler. Under this method, samples of the vector $\boldsymbol{\delta} = (\boldsymbol{\mu}', \dot{\boldsymbol{\lambda}}')'$ of small area parameters $\boldsymbol{\mu}$ and the model parameters $\dot{\boldsymbol{\lambda}}$ are partitioned, say $(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_r)$. Considering model (1.6), we have $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)'$ and $\dot{\boldsymbol{\lambda}} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)'$, the partitions can be $\boldsymbol{\delta}_1 = \boldsymbol{\beta}$, $\delta_2 = \mu_1, \ldots, \delta_{A+1} = \mu_A$, $\delta_{A+2} = \sigma_v^2$ and $\delta_{A+3} = \sigma_e^2$ so that in this case $r = A + 3$. To construct the Markov chain, a one-step transition probability or transition kernel $P(\cdot|\cdot)$ is specified such that the stationary distribution of the chain generated by $P(\cdot|\cdot)$ is the joint posterior density $f(\boldsymbol{\delta}|\boldsymbol{y})$. In other words, the chain $\{\boldsymbol{\delta}^j, j = 0, 1, 2, \ldots\}$ is constructed such that the distribution of $\{\boldsymbol{\delta}^j\}$ converges to a unique stationary (invariant) distribution equal to $f(\boldsymbol{\delta}|\boldsymbol{y})$. This is done to avoid the difficulty brought about by the intractable denominator in the joint posterior density of $f(\boldsymbol{\delta}|\boldsymbol{y})$, see Carlin and Louis (2008) for detailed explanation.

The Gibbs sampler constructs $P(\cdot|\cdot)$ by formulating a set of conditional distributions of the partitioned vector of parameters: $f(\boldsymbol{\delta}_1|\boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_r, \boldsymbol{y})$, $f(\boldsymbol{\delta}_2|\boldsymbol{\delta}_1, \boldsymbol{\delta}_3, \ldots, \boldsymbol{\delta}_r, \boldsymbol{y})$, …, $f(\boldsymbol{\delta}_r|\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{r-1}, \boldsymbol{y})$. The product of these conditional distributions is the transition kernel. If the conditional distributions have a standard form such as normal and gamma distributions, then samples can directly be generated from them. Otherwise, other methods can be used, for example the Metropolis-Hastings (M-H) algorithm, see Siddhartha and Greenberg (1995) for detailed exposition on the M-H method.

The Gibbs sampling algorithm involves three steps: (1) specification of the initial values of the parameter vector $\boldsymbol{\delta}^0$, initial values could be arbitrary. (2) generation of the chain: $\boldsymbol{\delta}^{(j+1)} = (\boldsymbol{\delta}_1^{(j+1)}, \ldots, \boldsymbol{\delta}_r^{(j+1)})$ by drawing $\boldsymbol{\delta}_1^{(j+1)}$ from $f(\boldsymbol{\delta}_1|\boldsymbol{\delta}_2^{(j)}, \ldots, \boldsymbol{\delta}_r^{(j)}, \boldsymbol{y})$; $\boldsymbol{\delta}_2^{(j+1)}$ from $f(\boldsymbol{\delta}_2|\boldsymbol{\delta}_1^{(j+1)}, \boldsymbol{\delta}_3^{(j)}, \ldots, \boldsymbol{\delta}_r^{(j)}, \boldsymbol{y})$; ...; $\boldsymbol{\delta}_r^{(j+1)}$ from $f(\boldsymbol{\delta}_r|\boldsymbol{\delta}_1^{(j+1)}, \ldots, \boldsymbol{\delta}_{r-1}^{(j+1)}, \boldsymbol{y})$. (3) set $j = j + 1$ and repeat step 2. Similar to other sampling algorithms, this method generates correlated samples, to overcome this problem, systematic samples from the simulations are usually taken to reduce the serial dependence in the samples. In addition, 'burn in' samples are discarded to remove those samples from the chain which occur before the chain has converged.

Details of various applications of the HB method to small area estimation are presented by Rao (2003). Here, we will only provide an overview of the application of the HB method to the estimation of small area parameters in the context of the linear mixed model in (1.7) with equal error variances (i.e., $k_{ah} = 1$). Assuming the population size N_a is large or that the sampling rate is negligible, we can take the ath small area mean as $\mu_a = \bar{\boldsymbol{X}}_a\boldsymbol{\beta} + \upsilon_a$. To generate the HB estimate of the parameters μ_a, first we express model (1.7) in HB form assuming that the sample data y_{ah} and $\mathbf{x}_{ah}$ obeys the said population model:

$$\begin{aligned}
&\text{(i)} \quad y_{ah}|\boldsymbol{\beta}, \upsilon_a, \sigma_e^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_{ah}\boldsymbol{\beta} + \upsilon_a, \sigma_e^2), \qquad h = 1, \ldots, n_a; \quad a = 1, \ldots, A \\
&\text{(ii)} \quad \upsilon_a|\sigma_\upsilon^2 \overset{\text{ind}}{\sim} N(0, \sigma_\upsilon^2), \qquad a = 1, \ldots, A \\
&\text{(iii)} \quad f(\boldsymbol{\beta}, \sigma_\upsilon^2\sigma_e^2) = f(\boldsymbol{\beta})f(\sigma_\upsilon^2)f(\sigma_e^2) \propto f(\sigma_\upsilon^2)f(\sigma_e^2)
\end{aligned} \tag{1.19}$$

Assuming an uninformative (or flat) prior for $\boldsymbol{\beta}$, i.e., $f(\boldsymbol{\beta}) \propto 1$ and gamma priors on σ_υ^{-2} and σ_e^{-2}, i.e., $\sigma_\upsilon^{-2} \sim G(a_\upsilon, b_\upsilon)$, $a_\upsilon \geq 0$, $b_\upsilon > 0$ and $\sigma_e^{-2} \sim G(a_e, b_e)$, $a_e \geq 0$, $b_e > 0$, Rao (2003) derived the following Gibbs conditionals:

$$(1) \quad [\boldsymbol{\beta}|\boldsymbol{\upsilon}, \sigma_\upsilon^2, \sigma_e^2, \boldsymbol{y}] \sim N[(\sum_a\sum_h \mathbf{x}'_{ah}\mathbf{x}_{ah})^{-1}\sum_a\sum_h \mathbf{x}'_{ah}(y_{ah} - \upsilon_a), \sigma_e^2(\sum_a\sum_h \mathbf{x}'_{ah}\mathbf{x}_{ah})^{-1}]$$

$$(2) \quad [\upsilon_a|\boldsymbol{\beta}, \sigma_\upsilon^2, \sigma_e^2, \boldsymbol{y}] \sim N[\gamma_a(\bar{y}_a - \bar{\mathbf{x}}_a\boldsymbol{\beta}), \gamma_a\frac{\sigma_e^2}{n_a}]$$

$$(3) \quad [\sigma_e^{-2}|\boldsymbol{\beta}, \boldsymbol{\upsilon}, \sigma_\upsilon^2, \boldsymbol{y}] \sim G[\frac{n}{2} + a_e, \frac{1}{2}\sum_a\sum_h(y_{ah} - \mathbf{x}_{ah}\boldsymbol{\beta} - \upsilon_a)^2 + b_e]$$

$$(4) \quad [\sigma_\upsilon^{-2}|\boldsymbol{\beta}, \boldsymbol{\upsilon}, \sigma_e^2, \boldsymbol{y}] \sim G(\frac{A}{2} + a_\upsilon, \frac{1}{2}\sum_a \upsilon_a^2 + b_\upsilon)$$

where $n = \sum_a n_a$, $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_A)'$, $\bar{y}_a = \sum_h y_{ah}/n_a$, $\bar{\mathbf{x}}_a = \sum_h \mathbf{x}_{ah}/n_a$, and $\gamma_a = \sigma_\upsilon^2/(\sigma_\upsilon^2 + \sigma_e^2/n_a)$. MCMC samples $\{\boldsymbol{\beta}^{(j)}, \boldsymbol{\upsilon}^{(j)}, \sigma_\upsilon^{2(j)}, \sigma_e^{2(j)}, j = \dot{d} + 1, \ldots, \dot{d} + \dot{D}\}$ can be directly derived from the conditional distributions above. The notation $\dot{d}$ here is the set of initial iterations (burn in) that are discarded in order to ensure that the remaining samples are drawn from a distribution close enough to the true stationary distribution (joint posterior density) to be useful for estimation and/or inference. The marginal samples $\{\boldsymbol{\beta}^{(j)}, \boldsymbol{\upsilon}^{(j)}\}$ can be used directly to estimate the posterior mean μ_a as

$$\hat{\mu}_a = \frac{1}{\dot{D}} \sum_{j=\dot{d}+1}^{\dot{d}+\dot{D}} \mu_a^{(j)} = \mu_a^{(\cdot)} \tag{1.20}$$

where $\mu_a^{(j)} = \bar{\mathbf{X}}_a \boldsymbol{\beta}^{(j)} + \upsilon_a^{(j)}$ and $(\cdot)$ in $\mu_a^{(\cdot)}$ signifies a mean . The corresponding posterior variance of μ_a is estimated as:

$$\hat{V}(\mu_a|\boldsymbol{y}) = \frac{1}{\dot{D}-1} \sum_{j=\dot{d}+1}^{\dot{d}+\dot{D}} (\mu_a^{(j)} - \mu_a^{(\cdot)})^2 \tag{1.21}$$

The HB method and EBLUP method have been compared by Rao (2003) using the classical data on county crop areas from the work of Battese et al. (1988). The small area estimates are close to each other, however the standard errors from HB are slightly larger reflecting the incorporation of additional uncertainty into the model.

1.5 Small Area Estimation Techniques Using Survey Weights

In modeling survey data, there is an issue regarding whether or not to account for the survey weights in the estimation procedure. Some statisticians have viewed weights as irrelevant while others consider them important and would generally incorporate the weights into every analysis. These survey weights are considered important since they reflect unequal inclusion probabilities and account for non-response and frame undercoverage in the survey. Accounting for the weights in the estimation process would mean generating design consistent estimators, i.e., the difference between the weighted estimates and the true value converges to zero in probability as the sample size grows large (Isaki and Fuller, 1982). In general, the local area sample size n_a is small, but some areas might have bigger sample sizes, in which case design consistency becomes relevant (You and Rao, 2002). Moreover, “weights can be used to protect against misspecification of the model holding in the population” (Pfefferman, 1993).

The survey weights w_{ah} are not considered in any of the estimation procedures discussed in the previous Section. Hence, the estimates generated will not generally be design consistent as the sample size in each small area or cluster increases. Moreover, as noted by Rao (2003), design-consistent model-based estimators are appealing because such estimators provide protection against model failures as the small area sample size increases. Survey-weighted small area estimators are discussed in this Section.

1.5.1 Pseudo-Estimated Best Linear Unbiased Prediction

You and Rao (2002) developed an estimation method called Pseudo-EBLUP (PEB) which extends the EBLUP method by considering the survey weights in the estimation process. The linear mixed model considered for PEB is a special case of (1.6) such that ($k_{ah} = 1$ for all (a,h)). Under this method, the unit level model in (1.6) is aggregated to obtain the survey-weighted area level model:

$$\bar{y}_{aw} = \bar{\mathbf{x}}_{aw}\boldsymbol{\beta} + v_a + \bar{e}_{aw} \tag{1.22}$$

with weights $\tilde{w}_{ah} = w_{ah}/\sum_h w_{ah} = w_{ah}/w_{a.}$, such that w_{ah} is the basic sampling weight. In the model above, $\bar{y}_{aw} = \sum_h \tilde{w}_{ah} y_{ah}$, $\bar{\mathbf{x}}_{aw} = \sum_h \tilde{w}_{ah}\mathbf{x}_{ah}$, $\bar{e}_{aw} = \sum_h \tilde{w}_{ah} e_{ah}$ such that $E(\bar{e}_{aw}) = 0$, $V(\bar{e}_{aw}) = \sigma_e^2 \sum_h \tilde{w}_{ah}^2 = \sigma_e^2/d_{a.(w)}$ and $d_{a.(w)} = \sum_h d_{ah(w)}$ where $d_{ah(w)} = \tilde{w}_{ah}^{-2}$.

As shown in Section 1.4.2 when the variance components σ_e^2 and σ_v^2 are assumed known, the BLUP of the small area mean is given by equation (1.12), however the method used for estimating the parameter $\boldsymbol{\beta}$ does not allow for the incorporation of the survey weights. Under the PEB method, survey weights are incorporated by deriving the estimate of $\boldsymbol{\beta}$ using a survey-weighted estimating equation. It involves obtaining the BLUP of v_a (given the parameters $\boldsymbol{\beta}$, σ_v^2, and σ_e^2 assumed known) from the aggregated area level model (1.22) as

$$\tilde{v}_{aw}(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2) = \gamma_{aw}(\bar{y}_{aw} - \bar{\mathbf{x}}_{aw}\boldsymbol{\beta}) \tag{1.23}$$

where $\gamma_{aw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/d_{a.(w)})$, then by solving the survey-weighted estimating equations:

$$\sum_{a=1}^{A}\sum_{h=1}^{n_a} w_{ah}\mathbf{x}_{ah}[y_{ah} - \mathbf{x}_{ah}\boldsymbol{\beta} - \tilde{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2)] = 0 \tag{1.24}$$

the estimator of $\boldsymbol{\beta}$ is obtained as:

$$\hat{\boldsymbol{\beta}}_w = \{\sum_{a=1}^{A}\sum_{h=1}^{n_a}\mathbf{x}_{ah}\kappa_{ah}\}^{-1}\{\sum_{a=1}^{A}\sum_{h=1}^{n_a}\kappa'_{ah}y_{ah}\} \tag{1.25}$$

where $\kappa_{ah} = w_{ah}(\mathbf{x}_{ah} - \gamma_{aw}\bar{\mathbf{x}}_{ah})$ and the corresponding covariance matrix as:

$$\begin{aligned}\boldsymbol{\Phi}_w =&\sigma_e^2(\sum_{a=1}^{A}\sum_{h=1}^{n_a}\mathbf{x}'_{ah}\kappa_{ah})^{-1}(\sum_{a=1}^{A}\sum_{h=1}^{n_a}\kappa'_{ah}\kappa_{ah})\{(\sum_{a=1}^{A}\sum_{h=1}^{n_a}\mathbf{x}'_{ah}\kappa_{ah})^{-1}\}' \\ &+ \sigma_v^2(\sum_{a=1}^{m}\sum_{h=1}^{n_a}\mathbf{x}'_{ah}\kappa_{ah})^{-1}\{\sum_{a=1}^{A}(\sum_{h=1}^{n_a}\kappa_{ah})'(\sum_{h=1}^{n_a}\kappa_{ah})\}\{(\sum_{a=1}^{A}\sum_{h=1}^{n_a}\mathbf{x}'_{ah}\kappa_{ah})^{-1}\}'\end{aligned} \tag{1.26}$$

which is similar to the sandwich estimator and is a robust variance estimator (Kauermann and Carroll, 2001). The estimators $\hat{\boldsymbol{\beta}}_w$ and $\boldsymbol{\Phi}_w$ depend on the variance components σ_v^2 and σ_e^2 which are replaced by the estimators given in equation (1.13) and (1.14), respectively. Then replacing the parameters in equation (1.12) with the corresponding estimators, we obtain the PEB estimators of the small area means which are claimed to satisfy the benchmarking property, i.e. when the estimator $\hat{\mu}_a$ is aggregated over the small areas it would be equal to the direct survey regression estimator of the overall total, assuming that the survey weights w_{ah} are calibrated to agree with the population total N_a (You and Rao, 2002).

1.5.2 Iterative Weighted Estimating Equation

In the description of the PEB estimator in the previous Section, the variance components σ_v^2 and σ_e^2 are estimated without using the survey weights since they were computed using the Henderson's method presented in Section 1.4.2. You et al. (2003) proposed an estimation procedure called Iterative Weighted Estimating Equation (IWEE) which allows for survey-weighted estimation of the parameter $\boldsymbol{\beta}$ and the variance components simultaneously. Similar to PEB, the IWEE method is based on the aggregated (using the survey weights) unit level model given in equation (1.22) and also uses the survey weighted estimating equation to obtain the estimate of the parameter $\boldsymbol{\beta}$. Hence, its estimator for $\boldsymbol{\beta}$ is similar to $\hat{\boldsymbol{\beta}}_w$ given in equation (1.25), as well as the covariance matrix in equation (1.26). However, You et al. (2003) derived the estimator of σ_e^2 and σ_v^2 as follows:

$$\hat{\sigma}_{ew}^{2(\dot{t})} = \frac{\sum_{a=1}^{A}\sum_{j=1}^{n_a} w_{ah}[y_{ah} - \bar{y}_{aw} - (\mathbf{x}_{ah} - \bar{\mathbf{x}}_{ah})\hat{\boldsymbol{\beta}}_w^{(\dot{t}-1)}]^2}{\sum_{a=1}^{A}[(1 - d_{a.(w)}^{-1})\sum_{h=1}^{n_a}\tilde{w}_{ah}]} \equiv \tilde{\sigma}_{ew}^{2(\dot{t})}(\boldsymbol{\beta}) \qquad (1.27)$$

and

$$\hat{\sigma}_{vw}^{2(\dot{t})} = \frac{1}{A}\sum_{a=1}^{A}\tilde{\upsilon}_{aw}^2 + \frac{\tilde{\sigma}_{vw}^{2(\dot{t}-1)}}{A}\sum_{a=1}^{A}(\gamma_{aw}-1)^2 + \frac{\tilde{\sigma}_{ew}^{2(\dot{t})}}{A}\sum_{a=1}^{A} d_{a.(w)}^{-1}\gamma_{aw}^2 \equiv \tilde{\sigma}_{vw}^{2(\dot{t})}(\tilde{\upsilon}_w, \sigma_e^2, \sigma_v^2) \quad (1.28)$$

The survey weighted estimates of $\boldsymbol{\beta}$, σ_e^2, σ_v^2 are obtained simultaneously by following iterative updating steps, where $\dot{t}$ in the equation above stands for the $\dot{t}$th iteration. Note that the notation for an iteration is a t with a dot on top in order to differentiate from t with no dot that is used in other Chapters to denote time period (e.g., t_0 to denote the census period). The variance component estimate based on the Henderson's method can be used as the initial value for the iterative steps. This approach is similar to the probability-weighted iterative generalized least squares (PWIGLS) method proposed by Pfefferman et al. (1998) for fitting multilevel models where the estimation process considered the unequal selection probabilities at each stage of sampling and adopted the iterative generalized least squares (IGLS) method. Under the IGLS method, the estimates are generated by iterating between the parameter $\boldsymbol{\beta}$ and the variance components until convergence (Goldstein, 2003).

1.5.3 Pseudo-HB method

An extension of the HB method analogous to the pseudo-EBLUP method is proposed by You and Rao (2003). This method is called Pseudo-HB (PHB) and like the PEB method it also satisfies the benchmarking property. Using the HB model specification in (1.19), the estimation method proceeds as follows: the samples $\{\sigma_e^{2(j)}, \sigma_v^{2(j)}\}$ from the unit level Gibbs sampler are obtained to calculate the survey-weighted estimator $\tilde{\boldsymbol{\beta}}_w^{(j)}$ in (1.25) as well as $\boldsymbol{\Phi}_w^{(j)}$ in (1.26). Then a new sample $\{\boldsymbol{\beta}_w^{(j)}\}$ is generated from $N(\tilde{\boldsymbol{\beta}}_w^{(j)}, \boldsymbol{\Phi}_w^{(j)})$. The set $\{\sigma_e^{2(j)}, \sigma_v^{2(j)}, \boldsymbol{\beta}_w^{(j)}\}$ are used to construct the pseudo-posterior mean and variance similar to (1.20) and (1.21), respectively. The PHB estimation procedure is a combination of the HB and PEB method. This method is also related to the model-based approach proposed by Pfefferman et al. (2006), which involves deriving the hierarchical model for a given sample data as a function of the population model and the selection probabilities, and then fitting the sample model using Bayesian approach by use of Markov chain Monte Carlo algorithm, e.g. the Gibbs sampler.

1.6 Summary

In this Chapter the GLMMs have been described as the framework for various small area estimation methods. Three different small area estimation techniques, and their corresponding modifications to incorporate survey weights in the estimation process, have been illustrated specifically for the linear mixed models, a special case of the GLMMs. The linear mixed model considered is what Rao (2003) called a one-fold nested error linear regression model. This model is fitted to a data set that has a multilevel, hierarchical, clustered or nested structure to generate estimates for the small areas of interest. Specifically, in terms of multi-level structure, the data can be considered to have a two-level structure such that the small areas (clusters) are the level 2 units and the subjects (e.g. households or individuals) within the small area are the level 1 units. This data structure is not necessarily similar to the data available for generation of small area estimates of poverty measures in Third World countries.

In most Third World countries where projects for small area estimation of poverty measures have been conducted, the structure of the survey data available is also clustered or nested. However, the small areas of interest are not necessarily the level 2 units mentioned above. The structure of the data usually has a three-level hierarchy - the small areas of interest are the level 3 (instead of level 2) units. The units in level 2 are usually the primary sampling units in the survey data and level 1 units are the households. Another level could be added if individuals within households are considered; the small areas could then be considered as level 4 units.

Given the three-level data structure and the income or consumption-based method of measuring poverty, a small area estimation procedure has been proposed by Elbers et al. (2003) which is discussed in detail in the next Chapter. This method involves a linear mixed model, specifically a one-fold nested error linear regression model (not accounting for the level 3 units or area level variation) for income. Predictions from this model are transformed in order to generate the required poverty measures and then aggregated up to the small area level of interest.

Chapter 2

Small Area Estimation of Poverty Measures

2.1 Introduction

Effective targeting of interventions and assistance aimed at alleviating or eradicating poverty requires reliable information on the poor and their location. Small area estimation is one statistical technique that can be used to generate the required information. At present the most widely implemented small area estimation technique for poverty measures in Third World countries is the method proposed by Elbers Lanjouw and Lanjouw (ELL) outlined in Elbers et al. (2003) and Elbers et al. (2002). A detailed review of the ELL method is presented in this Chapter and important features of the technique are compared with the "standard" small area estimation techniques that have been discussed in the previous Chapter.

This Chapter begins with an overview on measures of poverty (Section 2.2). This is followed by the presentation of the ELL method (Section 2.3) which includes the description of the ELL model, details of the parameter estimation technique and the generation of small area estimates. The ELL parameter estimation technique is compared with the other estimation techniques - PEB, IWEE and the general survey regression (GSR) method in Section 2.4 which includes an application of the different estimation techniques to real data from the Philippines. A summary of the Chapter is presented in the last Section (Section 2.5).

2.2 Measures of Poverty

Poverty is a very complex multidimensional phenomenon: there exists a wide array of definitions and methods of measurement and even after several decades of research on poverty, there is no consensus among researchers. Generally, poverty measurement is purely economic or monetary-based. It involves the formulation of a poverty line or a standard such that if income or consumption of a household falls below the said standard its members are considered poor. There are two popular approaches for determining poverty lines: (1) defining a nutritional basket considered to be sufficient for the healthy survival of a typical family - *absolute poverty line* and (2) the standard or line is set arbitrarily in relation to the average expenditure or income in a country - *relative poverty line.* Absolute poverty lines are commonly used in Third World or poorer countries while relative poverty lines are used in wealthier countries as mentioned in the Asian Development Bank (ADB) online resources (ADB, 2006).

Traditionally, there are two common measures of poverty used in most countries: (1) the head count ratio which is also known as poverty incidence ($\mathbb{H}$) and strictly it is poverty prevalence, defined as the proportion of the population whose income or expenditure is below the established poverty line, i.e., $\mathbb{H} = Q/N$, where Q is the number of individuals or households whose income or expenditure falls below the poverty line and N is the size of the population; (2) the income gap ratio ($\mathbb{I}$) which measures the depth of poverty and gives information on average income levels or shortfalls below the poverty lines. This is computed as $\mathbb{I} = (1/Q)\sum_{Y_h<\ell}[(\ell - Y_h)/\ell]$, where ℓ is the poverty line which could vary depending on area within a particular country and Y_h is the income of individual or household h, here $h = 1, \ldots, N$. The poverty measures $\mathbb{H}$ and $\mathbb{I}$ are usually combined to form the poverty gap ratio of the total population: $\mathbb{P} = \mathbb{H} * \mathbb{I} = (1/N)\sum_{Y_h<\ell}[(z - Y_h)/\ell]$, such that the poverty gap ratio is zero for the non-poor population. This represents the per capita monetary amount needed to bring all poor individuals to a basic level. In Third World countries, another measure of poverty called poverty severity (PS) is also usually published by national statistical agencies which measures the average squared distance below the poverty line, thereby giving more weight to the very poor. These three measures can be placed in a common framework proposed by Foster et al. (1984), the so-called Foster-Greer-Thorbeck (FGT) measures:

$$P_\flat = \frac{1}{N}\sum_{h=1}^{N}\left(\frac{\ell - Y_h}{\ell}\right)^\flat \cdot I(Y_h < \ell) \tag{2.1}$$

where N, Y_h and ℓ are as defined above. $I(Y_h < \ell)$ is an indicator function (equal to 1 when income or expenditure is below the poverty line, and 0 otherwise). Poverty incidence, gap and severity correspond to $\flat$=0,1 and 2, respectively.

Poverty incidence and poverty gap are not sensitive to the distribution of poverty, i.e., these two measures do not account for the level of deprivation among poor people, whether the household or individual belongs to the poorest of the poor or the marginally poor they are equally treated or weighted. On the other hand, poverty severity is distribution-sensitive such that more weight is given to the poorest household or individual in a given area. Hence under the FGT measures if $\flat$ has a value of at least 2, then the FGT index generates a measure that is sensitive to poverty distribution (Osberg and Xu, 2008). The FGT index is a modification of the pioneering poverty measurement approach proposed by Sen (1976) that addressed the limitations of poverty incidence and poverty gap by incorporating a measure of inequality of income distribution of the poor into the poverty index. Sen's poverty measure is as follows:

$$S_{po} = \mathbb{H}[\mathbb{I} + (1 - \mathbb{I})G(Y_{po})]$$

where $G(Y_{po})$ is the Gini coefficient of income distribution of the poor, a measure of inequality or statistical dispersion proposed by an Italian statistician Corrado Gini (1912) (see Sen (1976) for detailed description of the coefficient). The term Y_{po} refers to income of the poor, while $\mathbb{H}$ and $\mathbb{I}$ are as defined above. Various extensions and modifications to Sen's poverty measure have been proposed, see for example Kakwani (1980), Blackorby and Donaldson (1980), Chakravarty (1983), Foster et al. (1984), Foster and Shorrocks (1991), and many others. All these methods are formulated in an attempt to satisfy the different axioms for poverty measures or in such a way that the poverty index generated has the important characteristics of a "standard" poverty measure proposed by Sen (1976) which were recently reviewed by Osberg and Xu (2008).

Monetary-based measures of poverty are accepted widely since their interpretation is simpler and hence easier to communicate to policymakers and other stakeholders. Poverty however is more than just an economic deprivation. Recently, an increasing number of researchers and institutions are adopting a more holistic approach to poverty estimation by incorporating into the estimation procedure the multi-dimensional aspects of poverty; an example is the work of Kanji and Chopra (2007). The downside of this approach however is that a new national survey needs to be administered which includes various questions covering different aspects assumed to affect the poverty status of individuals such as socio-economic activities of an individual, health, physical abilities (i.e. individual physical attributes), intellectual capabilities, among others.

On the other hand, there are also researchers advocating the principle that the poor themselves are the real experts of their situation. A poverty measure known as Participatory Poverty Index (PPI) is based on the opinion of poor people considering various factors deemed critical to their poverty situation. PPI is derived by conducting workshops or meetings in identified poor communities with the participation of community officials and poor households. In these meetings, participants are asked for their views on how to address the problems that give rise to their poverty and how to measure and monitor poverty in their area. For recently conducted workshops aimed to generate PPIs, see for example ADB (2001), ACTIONAID-International (2006) and Xiaoyun and Remenyi (2008).

In developing countries where the ELL method has been used for poverty mapping projects of the World Bank, the estimates of poverty measures generated are usually based on the FGT indices ($\flat = 0, 1, 2$), although some would also include Sen's index similar to the one conducted in Albania (Neri et al., 2005). In this Chapter, the ELL method is described based only on the three poverty measures under the FGT indices. In the proposed updating method for poverty measures discussed in Chapter 4 and 5, the method is illustrated for updating small area estimates of poverty incidence i.e. an FGT index such that $\flat = 0$.

2.3 SAE for Poverty Measures in Developing Countries

SAE techniques have recently been applied to generate local level estimates of poverty measures both in affluent and developing countries. The current focus of development and aid paradigm for Third World countries is poverty alleviation which necessitates local level information on poverty measures for aid allocation and monitoring. Hence there is a more extensive use of the SAE technique developed for poverty measures in Third World countries than in First World or more affluent countries. In the literature, one application of the SAE method to poverty measures is a project conducted by the National Research Council (NRC, 2000) in the United States. In this SAE project, the Fay-Herriot method (a special case of the EBLUP method described in the previous Chapter) was used to generate county-level estimates of the number of poor school-age children in the country. Data on the variable of interest was available in the Current Population Survey (CPS) but due to the small sample size (if available) in the counties, direct survey estimates are unreliable. Data on relevant auxiliary variables were taken from administrative records. Details of the method and data sets used for producing the county (small area) estimates are given by Rao (2003, chap. 7) and NRC (2000). The results of this project were used by the United States Department of Education to allocate funds to counties, and then the states distributed the allocated funds among school districts.

As noted earlier, the most widely implemented small area estimation methodology for poverty measures in developing countries is the method proposed by Elbers et al. (2003) commonly known as the ELL method. This small area estimation method has been used in almost all the poverty mapping projects of the World Bank conducted in collaboration with national statistical agencies in developing countries. In its original form, this method uses monetary-based poverty measures described in Section 2.2. The ELL method differs from the Fay-Herriot method used by the NRC in the United States in that the proposed ELL model is not directly fitted to the variable of interest - specific poverty measure. Under the ELL method a model (specifically a linear mixed model) is fitted to income/consumption at household level and then the FGT framework given in equation (2.1) is applied to generate the required poverty measures. This is just one of the deviations of the ELL method from the "standard" small area estimation method described in Chapter 1, other differences will be described in the next Sections.

In Third World countries, data on income/consumption is usually available in a national survey. In the usual application of standard small area estimation modeling, data on the variable of interest would be taken from the survey while data on the auxiliary variables would be taken from the census or administrative data and there is a one-to-one correspondence between the two sets of data either on individual/household or any other level of aggregation suitable for the modeling procedure. That is, the individual or household level income/consumption data from the survey corresponds to individual or household level auxiliary data in the census. However, due to confidentiality limitations in Third World countries, this is not generally possible. The solution adopted under the ELL method is to limit the individual/household level auxiliary variables to those available and having a consistent definition in both the survey and the census. Hence, the preliminary step of the ELL method is the identification of individual/household level auxiliary variables that are defined and measured consistently in both data sources. These are supplemented by auxiliary variables available in the census at higher aggregation levels such as census means available at small area level (e.g. municipality) or cluster level (e.g., barangay). In some countries additional auxiliary variables are also taken from a Geographic Information System (GIS) database.

The linear mixed model for income/consumption presented in Section 2.3.1 is then fitted using the values of the individual/household level auxiliary variables from the survey along with the census means and GIS data. The ELL model fitting technique is discussed in Section 2.3.2. The fitted model is then applied to the whole census data - the individual/household auxiliary variables and the census means together with the GIS data to generate predicted household incomes using the bootstrap method (details in Section 2.3.3). Then the FGT framework is applied to generate the required small area poverty measures (i.e., poverty incidence, gap and severity).

2.3.1 The ELL Model

The ELL method is implemented by fitting the following income/consumption model:

$$Y_{\dot{b}h} = \mathbf{X}_{\dot{b}h}\boldsymbol{\beta} + u_{\dot{b}h} \tag{2.2}$$

where $\dot{b} = 1, \ldots, \dot{B}$, $h = 1, \ldots, N_{\dot{b}}$, and $Y_{\dot{b}h}$ is the log-transformed income or expenditure of the hth unit or household in the $\dot{b}$th cluster; $\dot{B}$ is the total number of clusters under study and $N_{\dot{b}}$ is the total number of households or individuals in the $\dot{b}$th cluster. $\mathbf{X}_{\dot{b}h}$ is the set of auxiliary variables available in the survey and the census and as mentioned above this set of auxiliary variables is supplemented by either census means data or GIS data. u is the random error term representing that part of $Y_{\dot{b}h}$ that cannot be explained by $\mathbf{X}_{\dot{b}h}$. Note that $\dot{b}$ and $\dot{B}$ is different from the b and B

notation that is used to denote categories of the variable of interest in Chapter 4. We also note that income and expenditure data have a skewed distribution, hence, transformation is done to make the data more symmetrical.

The households where data on income is taken are usually not independent, but have some kind of natural groupings or clusters. Households that are close to each other or those in the same cluster or (at a larger scale) small area tend to be similar in many respects. In the survey data, the clusters are usually the primary sampling units (PSUs) used for survey design. To account for the clustering of households, the random error term u could be assumed to have the following specification:

$$u_{\dot{b}h} = \upsilon_{\dot{b}} + e_{\dot{b}h}$$

where υ and e are independent of each other and uncorrelated with $\mathbf{X}_{\dot{b}h}$, $\upsilon_{\dot{b}}$ is the error term held in common by the $\dot{b}$th group or cluster (e.g. barangay for the Philippines) and $e_{\dot{b}h}$ is the household level error within the cluster. The importance of each term is measured by its respective variance, σ^2_υ and σ^2_e. There are different methods for estimating these variances which are shown in the next Sections. Note that the assumption that the error terms are uncorrelated with the explanatory variables $\mathbf{X}_{\dot{b}h}$ is a very important assumption in the linear regression framework. If $\mathbf{X}_{\dot{b}h}$ is correlated with the error terms, it means $\mathbf{X}_{\dot{b}h}$ is correlated with unmeasured variables that are influencing $Y_{\dot{b}h}$. Since we cannot eliminate their influence from the effect of $\mathbf{X}_{\dot{b}h}$ on $Y_{\dot{b}h}$, we will consistently over-estimate the regression coefficients and hence the parameter estimates are no longer unbiased.

Based on the specification of the error term (u) above, model (2.2) can then be written as

$$Y_{\dot{b}h} = \mathbf{X}_{\dot{b}h}\boldsymbol{\beta} + \upsilon_{\dot{b}} + e_{\dot{b}h} \tag{2.3}$$

this model is similar in form to the unit level model or the basic linear mixed model in (1.6) from Section 1.4. We note that the form of model (2.3) may be similar to model (1.6) but the group or aggregation level being referred to is different, e.g., Y_{ah} in model (1.6) refers to the hth household in the ath small area, while $y_{\dot{b}h}$ above refers to the hth household in the $\dot{b}$th cluster and the cluster being referred to here is typically much smaller than the small areas for which estimates are sought. For example in the Philippines, estimates are sought at the municipal level which is composed of several clusters or barangays, or a higher aggregation level than the level for which the ELL model is formulated. The ELL model therefore does not account for the variability among small areas which could affect the small area estimates generated and the associated standard error estimates.

To account for the between-area variation, the ELL model for income/consumption may be modified or improved by adding a small area effect to the model given in (2.3) so that we will have:

$$Y_{a\dot{b}h} = \mathbf{X}_{a\dot{b}h}\boldsymbol{\beta} + \upsilon_a + u_{a\dot{b}} + e_{a\dot{b}h} \tag{2.4}$$

with $a = 1, \ldots, A$, $\dot{b}$ and h are as defined above. This model is similar to the so called two-fold nested error regression model described in Stukel and Rao (1999). Under the two-fold nested error regression model the area level effects $\{\upsilon_a\}$, the cluster level effects $\{u_{a\dot{b}}\}$ and the unit level errors $\{e_{a\dot{b}h}\}$ are assumed to be mutually independent. In general it is also assumed that $e_{a\dot{b}h} = k_{a\dot{b}h}\tilde{e}_{a\dot{b}h}$, where $k_{a\dot{b}h}$ are known constants to allow for heteroscedasticity in the unit level errors $\tilde{e}_{a\dot{b}h}$. Similarly, the importance of each term could be measured by their respective variances, σ_υ^2, σ_u^2 and σ_e^2. The random components ($\{\upsilon_a\}$, $\{u_{a\dot{b}}\}$, and $\{e_{a\dot{b}h}\}$) are generally each assumed to be `iid` and normally distributed. These assumptions hold for sample data from a multistage cluster design such that sample clusters within small areas are self-weighting as well as sample households within sampled clusters, or a sampling design such that some of the auxiliary variables $\mathbf{X}_{a\dot{b}h}$ are used in the selection of the sample. This is not generally the case for survey data from developing countries used for modeling income/consumption since sample clusters are not necessarily self-weighting.

2.3.2 ELL Model Fitting Technique

As mentioned above, the preliminary step for fitting the ELL model is the selection of auxiliary variables. As in any other regression fitting exercise, the selection of auxiliary variables is one of the crucial steps in modeling. This is also the most time consuming stage of the implementation of the ELL method. A large amount of time is necessary to search for variables that match consistently in both the survey and the census in terms of definition and measurement (in general matching in terms of average as assessed via standard error of the survey estimates). Having identified the possible auxiliary variables: individual or household level auxiliary variables and variables at higher aggregation levels (cluster or small area level) from the census and GIS data, the ELL model presented in the previous Section (model 2.3) is fitted using what Elbers et al. (2002) referred to as "weighted generalized least squares (GLS)" procedure. Assuming that the sample data y_{b} and $\mathbf{x}_{bh}$ obey the population model in (2.3) and is viewed as a linear model with block-diagonal covariance matrix structure so that in matrix notation the model for the sample can be written as:

$$\boldsymbol{y}_{b} = \mathbf{x}_{b}\boldsymbol{\beta} + v_{b}\mathbf{1}_{n_a} + \boldsymbol{e}_{b} \tag{2.5}$$

with $V(\boldsymbol{y}) = \mathbf{V} = \text{diag}(\mathbf{V}_{b})$ denoting the block-diagonal covariance matrix.

The parameter estimation method under the ELL method claimed to be a weighted GLS is implemented as follows:

1) Generate an initial estimate of the parameter $\boldsymbol{\beta}$ through weighted least squares (using weights from the sampling design).

2) Decompose the residuals $(\hat{u}_{bh})$ from step 1 as follows: $\hat{u}_{bh} = \hat{u}_{b.} + (\hat{u}_{bh} - \hat{u}_{b.}) = \hat{v}_{b} + \hat{e}_{bh}$, here the subscript (.) indicates an average over that index.

3) Compute the estimated cluster level variance component as follows:

$$\hat{\sigma}_v^2 = max\left(\frac{\sum_{b} w_{b}(u_{b.} - u_{..})^2}{\sum_{b} w_{b}(1 - w_{b})} - \frac{\sum_{b} w_{b}(1 - w_{b})\hat{\tau}_{b}^2}{\sum_{b} w_{b}(1 - w_{b})}; 0\right) \tag{2.6}$$

where $\hat{\tau}_{b}^2 = \sum_h (e_{bh} - e_{b.})^2/(n_{b}(n_{b} - 1))$; $w_{b} = \sum_h w_{bh}/\sum_{b}\sum_h w_{bh}$, is the by-cluster transformed sampling weights which sum to one across clusters, w_{bh} is the re-scaled sampling weights which sum to the total sample size and n_{b} is the number of sample households or individuals in each cluster. Note that $\hat{\tau}_{b}^2$ is an estimate of the variance of $e_{b.}$. See Elbers et al. (2002) and Elbers et al. (2003) for details of the derivation.

4) Compute the estimated household level variance component. Elbers et al. (2003) suggested a heteroscedastic model-based computation which uses a logistic-type link function to bound the variance as follows:

$$\sigma_{e,bh}^2(\boldsymbol{z}_{bh}, \boldsymbol{\alpha}, \mathbb{A}, \mathbb{B}) = \left[\frac{\mathbb{A}\exp(\boldsymbol{z}_{bh}'\boldsymbol{\alpha}) + \mathbb{B}}{1 + \exp(\boldsymbol{z}_{bh}'\boldsymbol{\alpha})}\right] \tag{2.7}$$

where $\mathbb{A}$ and $\mathbb{B}$ are the upper and lower bounds respectively, estimated with the parameter vector $\boldsymbol{\alpha}$ using a standard pseudomaximum likelihood procedure (Elbers et al., 2003), and where $\boldsymbol{z}_{\dot{b}h}$ are auxiliary variables. The authors claim that if a minimum bound of zero and a maximum bound of $\mathbb{A}^* = (1.05)\max\{e^2_{\dot{b}h}\}$ is imposed, in general this would yield similar estimates of $\boldsymbol{\alpha}$. These restrictions allow one to estimate the simpler form

$$\ln\left[\frac{e^2_{\dot{b}h}}{\mathbb{A}^* - e^2_{\dot{b}h}}\right] = \boldsymbol{z}'_{\dot{b}h}\boldsymbol{\alpha} + r_{\dot{b}h} \tag{2.8}$$

where $r_{\dot{b}h}$ is an error term and the other variables are as defined earlier. In most of the World Bank poverty mapping projects, slight modifications are usually made, for example, adding a constant ς to $e^2_{\dot{b}h}$ in model (2.8).

By using model (2.8), and employing the delta method, $\hat{\sigma}^2_{e,\dot{b}h}$ is computed as:

$$\hat{\sigma}^2_{e,\dot{b}h} = \left[\frac{\mathbb{A}^* C_{\dot{b}h}}{1 + C_{\dot{b}h}}\right] + \frac{1}{2}\hat{\sigma}^2_r\left[\frac{\mathbb{A}^* C_{\dot{b}h}(1 - C_{\dot{b}h})}{(1 + C_{\dot{b}h})^3}\right] \tag{2.9}$$

where $C_{\dot{b}h} = \exp\{\boldsymbol{z}'_{\dot{b}h}\hat{\boldsymbol{\alpha}}\}$, and $\hat{\sigma}^2_r$ is the estimated variance of the residuals under model (2.8). Heteroscedastic modeling is conducted on the assumption that variation at the household level depends on some covariates.

As will be shown in the next Section, generation of estimated poverty measures and their corresponding standard errors involve simulation of the parameter estimates and residual terms. Distributions are specified for the regression parameter as well as the cluster level error and the standardized household level residuals. The standardized household level residuals are computed as follows:

$$\hat{e}^*_{\dot{b}h} = \frac{\hat{e}_{\dot{b}h}}{\hat{\sigma}_{e,\dot{b}h}} - \frac{1}{n}\left(\sum_{\dot{b}h}\frac{\hat{e}_{\dot{b}h}}{\hat{\sigma}_{e,\dot{b}h}}\right) \tag{2.10}$$

where n is the total number of observations, i.e. $\sum_{\dot{b}} n_{\dot{b}}$.

Based on the SAS program written by Zhao (2006) which is part of World Bank software developed for small area estimation of poverty measures called PovMap, if the household level errors are not heteroscedastic then a "direct" computation of the household level variance component which is now denoted as $(\hat{\sigma}^2_e)$ as opposed to the heteroscedastic model-based $(\hat{\sigma}^2_{e,\dot{b}h})$ is used. Direct computation involves using the difference between the estimated mean square error from the initial weighted least squares regression (fitted in step 1) and the computed estimate of the cluster level variance component $\hat{\sigma}^2_v$ from step 3. The implementation of the ELL method that is used for the

5) Generate the estimated block-diagonal covariance matrix $\hat{V}(\mathbf{y})$ with the following components, $\hat{\mathbf{V}}_{\dot{b}} = (\hat{\sigma}^2_{e,\dot{b}h}\mathbf{I}_{n_{\dot{b}}} + \hat{\sigma}^2_v \mathbf{1}_{n_{\dot{b}}}\mathbf{1}'_{n_{\dot{b}}})$, when the household level variance component is based on a heteroscedastic model, which is the usual practice. Here $\hat{\sigma}^2_v$ is the estimated cluster level variance, while $\hat{\sigma}^2_e$ is the estimated household level variance, $\mathbf{I}_{n_{\dot{b}}}$ is an identity matrix, $\mathbf{1}'_{n_{\dot{b}}} = (1...1)$ is a constant vector,

6) Compute the estimate of the parameter $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}}_{ELL} = \left(\sum_{\dot{b}=1}^{\dot{B}} \mathbf{x}'_{\dot{b}}\mathbf{W}_{\dot{b}}\hat{\mathbf{V}}_{\dot{b}}^{-1}\mathbf{x}_{\dot{b}}\right)^{-1} \left(\sum_{\dot{b}=1}^{\dot{B}} \mathbf{x}'_{\dot{b}}\mathbf{W}_{b}\hat{\mathbf{V}}_{\dot{b}}^{-1}\mathbf{y}_{\dot{b}}\right) \tag{2.11}$$

with the corresponding estimated variance-covariance matrix,

$$V(\hat{\boldsymbol{\beta}}_{ELL}) = \mathbf{D}\left[\left(\sum_{\dot{b}=1}^{\dot{B}} \mathbf{x}'_{\dot{b}}\mathbf{W}_{\dot{b}}\hat{\mathbf{V}}_{\dot{b}}^{-1}\mathbf{W}_{\dot{b}}\mathbf{x}_{\dot{b}}\right)^{-1}\right]\mathbf{D} = \hat{\boldsymbol{\Phi}} \tag{2.12}$$

where $\mathbf{D} = (\sum_{\dot{b}=1}^{\dot{B}} \mathbf{x}'_{\dot{b}}\mathbf{W}_{\dot{b}}\hat{\mathbf{V}}_{\dot{b}}^{-1}\mathbf{x}_{\dot{b}})^{-1}$, $\mathbf{x}_{\dot{b}} = (\mathbf{x}'_{\dot{b}1}, \ldots, \mathbf{x}'_{\dot{b}n_{\dot{b}}})'$; $\mathbf{y}_{\dot{b}} = (y_{\dot{b}1}, \ldots, y_{\dot{b}n_{\dot{b}}})'$; and $\mathbf{W}_{\dot{b}}$ is a diagonal matrix of sampling weights and note that $\mathbf{x}_{\dot{b}h}$ is a $1 \times \dot{p}$ vector and $\mathbf{x}_{\dot{b}}$ is $n_{\dot{b}} \times \dot{p}$ matrix.

Examining equation (2.11) above, this estimator is related to the generalized regression estimator (see Lohr (1999)). The way however in which the weight matrix $\mathbf{W}_{\dot{b}}$ enters the calculation is rather simplistic and does not appear to be sensible. The correct approach based on 'pseudomaximum likelihood' was outlined by Pfefferman et al. (1998) and involves splitting $\mathbf{x}'_{\dot{b}}\hat{\mathbf{V}}_{\dot{b}}^{-1}\mathbf{x}_{\dot{b}}$ into separate sums of squares and cross-product terms, and weighting each appropriately - if we write $\hat{\mathbf{V}}_{\dot{b}}^{-1} = c\mathbf{I}_{n_{\dot{b}}} + d\mathbf{1}_{n_{\dot{b}}}\mathbf{1}'_{n_{\dot{b}}}$ then the appropriate weighting is $c\mathbf{x}'_{\dot{b}}\mathbf{W}_{\dot{b}}\mathbf{x}_{\dot{b}} + d\mathbf{x}'_{\dot{b}}\mathbf{W}_{\dot{b}}\mathbf{1}_{n_{\dot{b}}}\mathbf{1}'_{n_{\dot{b}}}\mathbf{W}_{\dot{b}}\mathbf{x}_{\dot{b}}$.

Moreover $\mathbf{W}_{\dot{b}}\hat{\mathbf{V}}_{\dot{b}}^{-1}$, is not generally symmetric, so neither is $\mathbf{D}$ in equation (2.12). As a consequence the supposed covariance matrix of $\hat{\boldsymbol{\beta}}_{ELL}$, $V(\hat{\boldsymbol{\beta}}_{ELL})$, is also not symmetric. The PovMap software developed for the ELL methodology solves this problem by taking the average of their $V(\hat{\boldsymbol{\beta}}_{ELL})$ and its transpose, thereby forcing the matrix to be symmetric. The correct covariance matrix for their estimator simply replaces the final $\mathbf{D}$ in equation (2.12) by its transpose (Haslett et al., 2010).

2.3.3 Generation of Small Area Estimates of Poverty Measures

After the regression model has been fitted to the survey data, it is then applied to the census data as a predictor at household level, i.e., the regression equation is used to find the predicted value for each census household per capita income or expenditure and is generated via

$$\hat{Y}_{\dot{b}h} = \mathbf{X}_{\dot{b}h}\hat{\boldsymbol{\beta}} \tag{2.13}$$

Here $\mathbf{X}_{\dot{b}h}$ are auxiliary variables from the census. Then the poverty measures of interest which are nonlinear functions of $\hat{Y}_{\dot{b}h}$ are generated by using a bootstrap procedure. Bootstrapping is a set of statistical techniques that use computer generated random numbers to simulate the distribution of an estimator (Efron and Tibshirani, 1993). A more detailed discussion of the bootstrap method is given in Chapter 6.

Bootstrapping involves the generation of several predicted values for income/consumption, i.e.,

$$\hat{Y}^{\dot{s}}_{\dot{b}h} = \mathbf{X}_{\dot{b}h}\hat{\boldsymbol{\beta}}^{\dot{s}} + \hat{v}^{\dot{s}}_{\dot{b}} + \hat{e}^{\dot{s}}_{\dot{b}h} \tag{2.14}$$

where $\dot{s} = 1, ..., \dot{S}$, $\dot{S}$ being the total number of independent random draws. An example of the simulation method under the ELL method is the bootstrap procedure implemented by Haslett and Jones (2005) such that $\hat{\boldsymbol{\beta}}^{\dot{s}}$ is drawn independently from a multivariate normal distribution with mean $\hat{\boldsymbol{\beta}}$ and variance $V(\hat{\boldsymbol{\beta}})$. For the cluster level effects, each $\hat{v}^{\dot{s}}_{\dot{b}}$ is taken from the empirical distribution of $v_{\dot{b}}$. For the household level effects, predicted values are generated by using the heteroscedastic model formulated in the estimation stage. First, $\hat{\boldsymbol{\alpha}}^{\dot{s}}$ is drawn from a multivariate normal distribution with mean $\hat{\boldsymbol{\alpha}}$ and variance $V(\hat{\boldsymbol{\alpha}})$, then it is combined with $\mathbf{z}_{\dot{b}h}$ to generate the predicted variance to be used for adjusting the household-level effect $\hat{e}^{\dot{s}}_{\dot{b}h} = \hat{e}^{*\dot{s}}_{\dot{b}h} \times \hat{\sigma}^{\dot{s}}_{e,\dot{b}h}$ where $\hat{e}^{*\dot{s}}_{bh}$ represents a random draw from the empirical distribution of $e^{*}_{\dot{b}h}$, either for the whole data set or just within the cluster chosen for $v_{\dot{b}}$. Note that these residual estimates $\hat{e}^{*}_{\dot{b}h}$ have been mean corrected to sum to zero (see equation (2.10)).

Each complete set of bootstrap values $\hat{Y}^{\dot{s}}_{\dot{b}h}$, for a fixed value of $\dot{S}$, will generate a set of small area estimates. In the case of income/expenditure based poverty estimates, each estimate of Y is exponentiated to give the predicted expenditure or income, say $\hat{E}_{\dot{b}h} = exp(\hat{Y}_{\dot{b}h})$, then equation (2.1) is applied to generate an estimate of the poverty measure $P_{\flat}$. The mean and standard deviation of a particular small area estimate, across all $\dot{S}$ values, then yields a point estimate and its standard error for that area. This simulation approach is the most widely implemented under the ELL method. Elbers et al. (2003) however also suggested an alternative method wherein the contribution of model uncertainty to the variance of the estimates is estimated using the delta method.

2.4 ELL and Other Parameter Estimation Methods

Given the limitations of the ELL method discussed in Section 2.3.2, there are alternative estimation procedure that can be considered for estimating the regression parameter $\boldsymbol{\beta}$ and its corresponding variance-covariance matrix. Three alternative techniques are presented here, namely Pseudo-EBLUP, IWEE and General Survey Regression (GSR) which are also discussed and presented in Haslett et al. (2010). Pseudo-EBLUP and IWEE have been described in Section 1.5, however these two estimation procedures will be applied under the framework in which the ELL method is used, i.e. the basic linear mixed model described in Section 2.3.1, such that the variability among small areas is not accounted for. The level of aggregation at which the model is formulated is at the cluster or primary sampling unit (PSU) level which is usually smaller than the small area of interest. The Pseudo-EBLUP and IWEE method have never been used in poverty mapping or regression parameter estimation involving variables from a survey data with complex design, while the GSR method (Lohr, 1999) was used in the implementation of the poverty mapping project in Bangladesh (Haslett and Jones, 2004), Philippines (Haslett and Jones, 2005) and Nepal (Haslett and Jones, 2006). The use of GSR as an estimation procedure is just one of the departures (or modifications) that can be made in the implementation of the ELL methodology in developing countries.

2.4.1 The Pseudo-EBLUP and IWEE Method

There is very limited published literature on the application to real data sets of the Pseudo-EBLUP and IWEE method. The existing publications are considering the clusters as the small area and often use the data in Battese et al. (1988) which contains information on hectares of corn and soybeans per segment for counties in North Central Iowa and is assuming simple random sampling within areas or clusters (i.e., self-weighting). An exception is the recent work of Militino et al. (2006), an application of Pseudo-EBLUP in the estimation of total area occupied by olive trees in Navarra, Spain in which the units are also self-weighting within clusters. For poverty estimation however the survey data under consideration cannot generally be regarded as self-weighting because they have been obtained from a complex survey design involving stratification and cluster sampling with unequal inclusion probabilities. Hence, Pseudo-EBLUP and IWEE techniques under the poverty estimation framework is applied in a more complex situation where the clusters (e.g., barangay) are different from the small area (e.g., municipality) and the cluster are sub-units of the small area and the sampling scheme is not self-weighting.

The Pseudo-EBLUP estimation has been described in Chapter 1 however as pointed out above, the income/consumption model is not entirely the same as the basic linear mixed model presented and described in Section 1.3 which is generally used in standard small area estimation applications. To differentiate the Pseudo-EBLUP parameter estimators from the ones described in Section 1.5.1, and for consistency with the income/consumption model, the estimator of $\boldsymbol{\beta}$ is now as follows:

$$\hat{\boldsymbol{\beta}}_w = \left\{ \sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \mathbf{x}_{\dot{b}h} \boldsymbol{\kappa}'_{\dot{b}h} \right\}^{-1} \left\{ \sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \boldsymbol{\kappa}_{\dot{b}h} y_{\dot{b}h} \right\} \tag{2.15}$$

where $\boldsymbol{\kappa}_{\dot{b}h} = w_{\dot{b}h}(\mathbf{x}_{\dot{b}h} - \gamma_{\dot{b}w}\bar{\mathbf{x}}_{\dot{b}h})$. The corresponding estimated covariance matrix is:

$$\begin{aligned} \boldsymbol{\Phi}_w =& \sigma_e^2 \left(\sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \mathbf{x}_{\dot{b}h} \boldsymbol{\kappa}'_{\dot{b}h} \right)^{-1} \left(\sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \boldsymbol{\kappa}_{\dot{b}h} \boldsymbol{\kappa}'_{\dot{b}h} \right) \left[\left(\sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \mathbf{x}_{\dot{b}h} \boldsymbol{\kappa}'_{\dot{b}h} \right)^{-1} \right]' \\ & + \sigma_v^2 \left(\sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \mathbf{x}_{\dot{b}h} \boldsymbol{\kappa}'_{\dot{b}h} \right)^{-1} \left[\sum_{\dot{b}=1}^{\dot{B}} \left(\sum_{h=1}^{n_{\dot{b}}} \boldsymbol{\kappa}_{\dot{b}h} \right) \left(\sum_{\dot{b}=1}^{n_{\dot{b}}} \boldsymbol{\kappa}_{\dot{b}h} \right)' \right] \left[\left(\sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} \mathbf{x}_{\dot{b}h} \boldsymbol{\kappa}'_{\dot{b}h} \right)^{-1} \right]' \end{aligned} \tag{2.16}$$

As emphasized in Section 1.5.1, the estimators $\hat{\boldsymbol{\beta}}_w$ and $\boldsymbol{\Phi}_w$ depend on the variance components σ_e^2 and σ_v^2 which are replaced by their corresponding estimates. Note that all the other notations not redefined are assumed to be the same as in Section 1.5.1.

The IWEE regression parameter estimator is similar to the Pseudo-EBLUP estimator above, as well as its covariance matrix estimator. IWEE differs from the Pseudo-EBLUP in the computation of the variance components (sampling weights are incorporated) and the manner by which the estimate of $\boldsymbol{\beta}$ and the variance components are generated as discussed in Section 1.5.2. The variance components equations under IWEE are now as follows:

$$\hat{\sigma}_{ew}^{2(\dot{t})} = \frac{\sum_{\dot{b}=1}^{\dot{B}} \sum_{h=1}^{n_{\dot{b}}} w_{\dot{b}h} [y_{ah} - \bar{y}_{\dot{b}w} - (\mathbf{x}_{\dot{b}h} - \bar{\mathbf{x}}_{\dot{b}h}) \hat{\boldsymbol{\beta}}_w^{(\dot{t}-1)}]^2}{\sum_{\dot{b}=1}^{\dot{B}} [(1 - d_{\dot{b}.(w)}^{-1}) \sum_{h=1}^{n_{\dot{b}}} \tilde{w}_{\dot{b}h}]} \equiv \tilde{\sigma}_{ew}^{2(\dot{t})}(\boldsymbol{\beta}) \tag{2.17}$$

and

$$\hat{\sigma}_{vw}^{2(\dot{t})} = \frac{1}{\dot{B}} \sum_{\dot{b}=1}^{\dot{B}} \tilde{\upsilon}_{\dot{b}w}^2 + \frac{\tilde{\sigma}_{vw}^{2(\dot{t}-1)}}{\dot{B}} \sum_{\dot{b}=1}^{\dot{B}} (\gamma_{bw} - 1)^2 + \frac{\tilde{\sigma}_{ew}^{2(\dot{t})}}{\dot{B}} \sum_{\dot{b}=1}^{\dot{B}} d_{\dot{b}.(w)}^{-1} \gamma_{\dot{b}w}^2 \equiv \tilde{\sigma}_{vw}^{2(\dot{t})}(\tilde{\upsilon}_w, \sigma_e^2, \sigma_v^2) \tag{2.18}$$

Note that the equations above are the same as those presented in Chapter one. The only difference is in the notation used for the PSUs which is now denoted as $\dot{b}$ instead of a.

2.4.2 The GSR Method

Another approach to generate the estimator of the parameter $\boldsymbol{\beta}$ and its associated covariance matrix is the GSR method which can be considered as an "aggregated approach" (Skinner et al., 1989) to analyze survey data such that the parameter of interest is defined unconditionally across clusters or population subgroups. The PEB and IWEE on the other hand are considered to belong to the "disaggregated approach" where parameters are defined conditionally on the population structure or survey design variables. The regression parameter estimate ($\boldsymbol{\beta}$) under the GSR given below is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population.

$$\hat{\boldsymbol{\beta}}_S = (\mathbf{x}'\mathbf{W}\mathbf{x})^{-1}\mathbf{x}'\mathbf{W}\mathbf{y} \tag{2.19}$$

This estimator is not derived under the model specified by (2.3) even under the homoscedastic variances for household level errors. The linearized/robust variance estimate for $\hat{\boldsymbol{\beta}}_S$ is based on the design-based variance estimator for a total, given as, as,

$$\hat{V}(\hat{\boldsymbol{\beta}}_S) = \mathbb{D}\left\{\frac{\dot{B}}{\dot{B}-1}\sum_{\dot{b}=1}^{\dot{B}}\left(\sum_{h=1}^{n_{\dot{b}}} w_{\dot{b}h}\mathbf{d}_{\dot{b}h}\right)'\left(\sum_{h=1}^{n_{\dot{b}}} w_{\dot{b}h}\mathbf{d}_{\dot{b}h}\right)\right\}\mathbb{D} \tag{2.20}$$

where $\mathbf{d}_{\dot{b}h} = e_{\dot{b}h}\mathbf{x}_{\dot{b}h}$; $e_{\dot{b}h}$, is the residual; $\mathbf{x}_{\dot{b}h}$ is a vector of the independent variables; $w_{\dot{b}h}$ is the sampling weight; $\mathbb{D} = (\mathbf{x}'\mathbf{W}\mathbf{x})^{-1}$; and $\mathbf{W}$ is a diagonal matrix of the sampling weights.

It can be observed from equation (2.19) above that the selection effects (survey weights) are incorporated in the estimation process but not the design variables such as the cluster effects and hence under the GSR method variance components associated with the cluster and household level effects are not needed for generating the regression parameter estimate and its associated estimated covariance matrix.

2.4.3 Comparison of the Parameter Estimation Methods

The ELL methodology is claimed to follow a "weighted GLS" estimation procedure, weighted in the sense that the method incorporates the survey weights in estimating the model parameters. The covariance structure of the error however is unknown, hence the implementation of the method basically follows a two-step Iteratively Reweighted Least Squares (IRLS) estimation procedure. In the first iteration the method uses a weighted least squares (using survey weights) regression and the residuals generated are used to improve the estimate of the covariance structure. As pointed out in Section 2.3.2, the sampling weights are not properly incorporated in the estimation process, leading to non-interpretability of the elements in some matrices involved in the estimation of the regression parameters as well as asymmetry in the covariance matrix.

The covariance structure is generated by estimating the cluster and household level variance components. Under this method, the procedure for the estimation of the variance components incorporates the sampling weights at the cluster level but not at the household level. The two ways (direct computation and heteroscedastic model-based) of estimating the household level variance components generate un-weighted estimates. Under the direct computation method, the household level variance component is determined from the residuals of the weighted least squares regression at the preliminary step and the weighted estimate of the cluster level component. The heteroscedasticity based computation is based on modeling the square of the residuals from the weighted least squares.

The implementation of the pseudo-EBLUP and IWEE methods follow an estimating function approach (Godambe, 1991) specifically, a survey-weighted estimating equation related to the the pseudomaximum likelihood approach described by Pfefferman et al. (1998). The pseudomaximum likelihood approach is similar to the maximum likelihood approach however the survey weights are incorporated in the formulation of the likelihood equation. These two procedures (Pseudo-EBLUP and IWEE) incorporate the sampling weights properly in the estimation of the regression parameter $\boldsymbol{\beta}$ and the corresponding standard error. However, the Pseudo-EBLUP method uses the Henderson's method in the estimation of the variance components which generates un-weighted estimates of the variance components. The IWEE method, on the other hand, incorporates the sampling weights iteratively from the estimation of variance components (cluster and household level) to the estimation of the parameter $\boldsymbol{\beta}$, as well as for the computation of the corresponding standard error.

The GSR method is the least complicated estimation procedure as it employs a weighted least square procedure and uses the sandwich estimator for estimating the variance of the regression parameter. This method however does not model the population structure and hence, generates the estimate of regression parameter and its corresponding standard error without computing the variance components. An advantage of this method however is that it is readily available in statistical packages such as STATA, Sudaan and WesVar.

The ELL method combines sampling weights and covariance structure in a way that is non-standard in that it produces an asymmetric estimated covariance matrix for the estimates of $\boldsymbol{\beta}$ and also uses this matrix in estimating $\boldsymbol{\beta}$ itself. For estimating $\boldsymbol{\beta}$ this would be acceptable if the asymmetric matrix were a generalized inverse of the correct covariance matrix. It is however clearly not acceptable as an estimated covariance matrix, a problem ELL attempt to circumvent (e.g., in the World Bank's PovMap software) by averaging each of the relevant pairs of off-diagonal elements to meet the necessary condition that a covariance matrix be symmetric as pointed out in Section 2.3.2.

Based on the discussion above, for all the techniques considered, the survey-based estimation procedure of the parameter $\boldsymbol{\beta}$ and its corresponding covariance matrix are theoretically sound given their assumptions, except for some inconsistencies in the estimation of the regression parameter and its associated covariance matrix under the ELL method. The IWEE is the method that best incorporates the sampling weights from the computation of the variance components necessary for the generation of small area estimates and their estimated standard errors. In terms of implementation, the GSR method would generally be the simplest option as it is available in statistical packages as mentioned above.

2.4.4 Application to Real Data

In this Section, the four different regression techniques (one of which contains two variants of ELL) are compared using the Philippine 2000 Family Income and Expenditure Survey (FIES) as in the Survey Methodology publication by Haslett et al. (2010). The FIES data is a nationwide survey undertaken by the Philippines National Statistics Office (NSO) every three years. The survey gathers details on family income and expenditure as well as information affecting income and expenditure. Selected households are interviewed in two separate operations, each covering a half-year period, in order to allow for seasonal patterns in income and expenditure. For FIES 2000 the interviews were conducted in July 2000, for the period 01 January to 30 June and January 2001 for the period 01 July to 31 December. The sample design for FIES used a multi-stage stratified random sampling technique. Barangays are the primary sampling units (PSUs) and are stratified into urban and rural within each province and selected using systematic sampling with probability proportional to size. Large barangays are further divided into enumeration areas and subjected to further sampling before the final stage in which households are systematically sampled from the 1995 Population Census List of Households. Interview non-response was only 3.4 percent, with 39,615 of the sample households being successfully interviewed in both survey visits.

The auxiliary variables used in this application are adopted from the variables included in the model formulated by Haslett and Jones (2005) that was fitted without using PovMap for the small area poverty mapping project in the Philippines. The auxiliary variables included both household characteristics and municipal means (in which the household data used have the same value for every sampled household in a given municipality, i.e., small area). These auxiliary variables are not only derived from the FIES data but also from the Philippine 2000 Labor Force Survey (LFS) and Census of Population and Housing (CPH). The LFS collects socioeconomic characteristics of the population over 15 years old. It is conducted on a quarterly basis by the NSO by personal interview, using the previous week as the reference period. Being part of the Integrated Survey of Households (NSCB, 2000), the July 2000 and January 2001 surveys used the same sample of households as the 2000 FIES. Thus the two data sets can be merged to form a richer set of auxiliary variables. Additional auxiliary variables were also taken from the 2000 CPH in the form of municipal means. Census variables in both the short and long form were averaged at municipal level to create new data sets that could be merged with the set of auxiliary variables from FIES and LFS.

The Regression Coefficients

Presented in Tables 2.1, 2.2, and 2.3 are the computed estimates of the parameter ($\boldsymbol{\beta}$) and the corresponding estimated standard errors as well as the estimates of the variance components at the national, regional and provincial levels, respectively. Table 2.2 shows one of the regional models of the 16 models fitted (see Appendix B) at the regional level (there were 16 regions in the Philippines in the year 2000). Similarly, Table 2.3 shows one of the provincial models of the 20 models formulated (see Appendix B) for 20 selected provinces. To standardize comparison, exactly the same set of predictor variables are used for all the different model fitting techniques. (There are five sets of parameter estimates, although there are only four basic methods considered, because ELL is used both with and without heteroscedasticity.) Note that in practice when ELL is applied, the survey data is often subdivided and separate models fitted to each subsample, e.g., to each regionally-based stratum as the 16 regions in the Philippines or even provincial level models. This can lead to overfitted models and downwardly biased estimated standard errors for small area estimates. For the analysis here, a single model (or the national level model) has also been fitted. In practice intermediate models with some but not all possible regional effects seem to work best, see for example Haslett and Jones (2005).

To assess the differences of the estimates generated from the different techniques, an informal comparison of the "significance" of the different estimates of β is conducted by subtracting from the estimate by one method the mean of the other methods' estimates, then dividing by the estimated standard error of the one method. At the national level (Table 2.1), estimates of the regression coefficients generated from the different methods are significantly different from each other for a number of the independent variables. GSR tends to generate estimates of the regression coefficients for the majority of the variables that are significantly different from the other methods. As pointed out earlier, the GSR estimator is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population and hence this estimator is not derived under the model specified by (2.3). However, it is the most conservative as it generates the highest estimated standard error for all the estimated regression coefficients of the household characteristics. On the other hand, the IWEE method has the highest estimated standard error for all the regression coefficients' estimates of the municipal means. The ELL_H (ELL with heteroscedasticity) method can be considered to be the least conservative since it produces the lowest estimated standard errors for all the regression coefficients' estimates of the household characteristics as well as for the municipal means, except for two variables where GSR generated the smallest estimates.

At the regional level (Table 2.2), estimates of the regression coefficients are generally similar for all the different estimation methods, except that the GSR and/or ELL_H methods generated estimates for a few variables which were significantly different from the other methods. Similar to the national level estimated standard errors, GSR also tends to be the most conservative method for the majority of the regional level models - it generated the highest estimated standard errors for most of the regression coefficients' estimates of the household characteristics. IWEE has the highest estimated standard error for most of the coefficients of the municipal means. The ELL_H method produces the lowest estimated standard errors for the majority of the estimated regression coefficients of the household characteristics and municipal means.

Similar to the regional level estimates, the regression coefficients' estimates at the provincial level are similar except for some discrepancies from the GSR and ELL_H estimates as shown in Table 2.3. For the estimated standard errors of the regression coefficients, the ELL_H still produces the lowest estimates for the majority of the coefficients of the household characteristics; however, the GSR method (instead of the ELL_H method) now produces the lowest estimated standard error for the majority of the coefficients for municipal means.

The Variance Components

For small area estimates of poverty measures, the estimated standard errors in the regression are only one part of the small area estimates' standard errors. There is also variation at the cluster level in (2.3) that needs to be considered (to different degrees depending on the level of aggregation used to construct the small areas) as well as variation at the household level. These additional sources of variation can be assessed via the estimated variance components.

At the national level, the ELL method generates the smallest estimated cluster level variance, which is about 92% of the Pseudo-EBLUP method and 86% of the IWEE method. As to the household level variance, the IWEE method generates the smallest estimate. A similar scenario applies to the estimates of the variance components at the regional level, the ELL method also tends to generate the smallest estimated cluster level variance with ratios to Pseudo-EBLUP and IWEE ranging from around 82% to 100%. The IWEE method still has the smallest estimated household level variance. The same situation holds at the provincial level, the ELL method once again

Table 2.1: National level estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1576633) **Based from the ELL method

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.11867	0.00181	-0.12034	0.00165	-0.11875	0.00183	-0.11888	0.00180	-0.11405	0.00216
famsizesqc	0.00937	0.00039	0.00981	0.00036	0.00938	0.00039	0.00939	0.00038	0.00898	0.00044
type_mult	0.03876	0.01697	0.03703	0.01588	0.03699	0.01717	0.03466	0.01692	0.11460	0.02194
per_kids	-0.20342	0.01476	-0.20818	0.01322	-0.20293	0.01491	-0.20216	0.01467	-0.22864	0.01617
roof_light	-0.06314	0.01291	-0.05808	0.01056	-0.06263	0.01306	-0.06175	0.01287	-0.09251	0.01413
per_61up	-0.09402	0.01420	-0.08331	0.01371	-0.09392	0.01435	-0.09389	0.01412	-0.09705	0.01698
roof_strong	0.05882	0.01135	0.05633	0.00962	0.05944	0.01148	0.06030	0.01132	0.03118	0.01293
wall_light	-0.05459	0.01182	-0.04979	0.00975	-0.05426	0.01195	-0.05392	0.01178	-0.06286	0.01353
wall_salvaged	-0.10814	0.02505	-0.11327	0.02058	-0.10748	0.02533	-0.10607	0.02495	-0.15702	0.02925
wall_strong	0.14248	0.01051	0.12964	0.00910	0.14274	0.01063	0.14319	0.01047	0.12662	0.01284
fa_xs	-0.17052	0.00941	-0.16756	0.00782	-0.17144	0.00952	-0.17236	0.00939	-0.14213	0.01110
fa_s	-0.08368	0.00861	-0.08242	0.00725	-0.08403	0.00871	-0.08454	0.00857	-0.06667	0.00964
fa_l	0.09016	0.00908	0.08478	0.00792	0.09065	0.00918	0.09106	0.00904	0.07848	0.01047
fa_xl	0.16959	0.01104	0.15404	0.00992	0.17034	0.01117	0.17121	0.01100	0.14300	0.01334
fa_xxl	0.27072	0.01144	0.24485	0.01094	0.27172	0.01157	0.27274	0.01140	0.23913	0.01457
fa_xxxl	0.36190	0.01371	0.31369	0.01286	0.36270	0.01387	0.36382	0.01367	0.32123	0.02025
all_eled	0.19084	0.01535	0.20497	0.01307	0.19031	0.01551	0.18964	0.01527	0.21344	0.01831
all_hsed	0.42325	0.01250	0.43771	0.01083	0.42192	0.01263	0.42024	0.01244	0.48180	0.01475
all_coed	1.21591	0.01371	1.29368	0.01379	1.21324	0.01386	1.20935	0.01366	1.35022	0.01827
dom_help	0.60207	0.01629	0.61218	0.01886	0.60035	0.01645	0.59733	0.01620	0.70307	0.02656
head_male	-0.05878	0.00988	-0.04581	0.00932	-0.05862	0.00998	-0.05819	0.00982	-0.07410	0.01173
no_spouse	-0.09367	0.00987	-0.07376	0.00917	-0.09361	0.00997	-0.09351	0.00981	-0.09599	0.01123
hou_9600	0.28537	0.07654	0.25643	0.07375	0.28871	0.07911	0.28783	0.08066	0.31956	0.07941
hea_rel_mus	0.09058	0.02645	0.10859	0.02507	0.09753	0.02728	0.09731	0.02782	0.10196	0.02737
Per_eng	0.17273	0.06529	0.14561	0.06298	0.17782	0.06754	0.17799	0.06887	0.17076	0.06407
Hou_coelpg	0.37463	0.04348	0.39784	0.04210	0.37934	0.04494	0.37792	0.04581	0.42682	0.03711
Hou_own_ref	0.17716	0.10497	0.18342	0.10178	0.17189	0.10843	0.17329	0.11055	0.13791	0.09766
Hou_own_tel	1.39287	0.13356	1.42109	0.12987	1.38551	0.13723	1.38974	0.13989	1.23506	0.13019
Per_wor_prh	0.46957	0.15484	0.40302	0.14926	0.47517	0.16006	0.47208	0.16317	0.50814	0.15210
Per_ind_52	-0.76245	0.21708	-0.78120	0.21073	-0.76326	0.22410	-0.76307	0.22849	-0.73294	0.21214
const	9.54013	0.05525	9.54456	0.05290	9.53566	0.05698	9.53594	0.05791	9.52622	0.05613
Variance										
Components	HH	Cluster	HH	Cluster	HH	Cluster	HH	Cluster	HH**	Cluster**
Estimate	level	level	level	level	level	level	level	level	level	level
	0.18461	0.04741	NA*	0.04741	0.18820	0.05172	0.18185	0.05498	0.18461	0.04741

Table 2.2: Regional level estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.18930) **Based from the ELL results

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.12327	0.00760	-0.12934	0.00689	-0.12377	0.00752	-0.12380	0.00749	-0.11786	0.00997
famsizesqc	0.01096	0.00164	0.01190	0.00147	0.01101	0.00163	0.01102	0.00162	0.01030	0.00195
dom_help	0.81037	0.08873	0.75624	0.10986	0.80727	0.08784	0.80708	0.08751	0.84490	0.08911
wall_light	-0.06808	0.04289	-0.06390	0.03743	-0.06020	0.04272	-0.05973	0.04257	-0.14472	0.04226
wall_strong	0.13761	0.03745	0.15212	0.03469	0.14514	0.03737	0.14560	0.03725	0.06116	0.04249
fa_xs	-0.22074	0.04910	-0.22368	0.04518	-0.22723	0.04875	-0.22761	0.04858	-0.14856	0.05665
fa_s	-0.13540	0.03840	-0.12255	0.03344	-0.13775	0.03805	-0.13789	0.03791	-0.11059	0.04538
fa_l	0.09484	0.03709	0.08894	0.03429	0.09590	0.03676	0.09597	0.03663	0.08529	0.04122
fa_xl	0.16627	0.04315	0.15519	0.04072	0.16938	0.04284	0.16958	0.04269	0.13698	0.04897
fa_xxl	0.33706	0.04545	0.31196	0.04829	0.34173	0.04516	0.34201	0.04500	0.29156	0.05148
fa_xxxl	0.33103	0.06185	0.30377	0.06029	0.33762	0.06134	0.33801	0.06111	0.26052	0.06635
all_hsed	0.33987	0.05253	0.35591	0.04783	0.33807	0.05209	0.33796	0.05189	0.35776	0.04843
all_coed	1.21824	0.05734	1.24762	0.05842	1.20787	0.05692	1.20726	0.05671	1.32979	0.06227
per_kids	-0.24699	0.06440	-0.24047	0.05846	-0.24439	0.06371	-0.24424	0.06347	-0.27423	0.07050
per_61up	-0.14609	0.06126	-0.15938	0.05787	-0.14703	0.06063	-0.14708	0.06040	-0.13525	0.07124
hou_9600	1.13985	0.49103	1.27035	0.47888	1.14320	0.52137	1.14357	0.52172	1.07509	0.51937
Hou_own_ref	1.45233	0.24550	1.51020	0.23864	1.44986	0.26072	1.44985	0.26089	1.44779	0.23585
const	9.36877	0.20322	9.32363	0.19660	9.36597	0.21502	9.36569	0.21512	9.41385	0.21430
Variance										
Components	HH	Cluster	HH	Cluster	HH	Cluster	HH	Cluster	HH**	Cluster**
Estimate	level	level	level	level	level	level	level	level	level	level
	0.19544	0.03073	NA*	0.03073	0.19052	0.03728	0.18902	0.03748	0.19544	0.03073

Table 2.3: Provincial level estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.23749) **Based from the ELL method

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1450	0.0175	-0.1489	0.0156	-0.1452	0.0179	-0.1449	0.0171	-0.1413	0.0097
famsizesqc	0.0090	0.0063	0.0124	0.0067	0.0091	0.0065	0.0090	0.0062	0.0085	0.0055
fa_xs	-0.4549	0.1126	-0.3816	0.1010	-0.4552	0.1149	-0.4546	0.1095	-0.4479	0.0718
fa_s	-0.2550	0.0976	-0.2653	0.0794	-0.2545	0.0995	-0.2555	0.0951	-0.2693	0.1198
wall_light	-0.2055	0.0945	-0.1474	0.0778	-0.2057	0.0965	-0.2058	0.0919	-0.2063	0.1070
all_hsed	0.4007	0.1643	0.3531	0.1448	0.4015	0.1673	0.4006	0.1601	0.3891	0.1585
all_coed	1.5411	0.1677	1.8202	0.1769	1.5429	0.1709	1.5429	0.1635	1.5439	0.2326
Hou_own_tel	3.4373	1.0270	3.2630	1.0582	3.4265	1.0622	3.4274	0.9871	3.4392	0.5733
Per_wor_prh	-1.1075	1.1933	-1.5801	1.2008	-1.1049	1.2327	-1.1056	1.1483	-1.1150	0.8729
const	10.0976	0.1480	10.0798	0.1279	10.0988	0.1517	10.0981	0.1435	10.0872	0.1373
Variance										
Components	HH	Cluster	HH	Cluster	HH	Cluster	HH	Cluster	HH**	Cluster**
Estimate	level	level	level	level	level	level	level	level	level	level
	0.25753	0.01871	NA*	0.25753	0.26682	0.02079	0.24498	0.01671	0.25753	0.01871

tends to generate the smallest estimated cluster level variance for most provinces with the smallest ratio to Pseudo-EBLUP about 53% and to IWEE about 48%. For a number of provinces, IWEE tends to generate the smallest estimated cluster level variance. For the household level variance, IWEE still generated the smallest estimate. Generally, estimates of the cluster level variance tend to be more variable at the provincial level which is due to smaller sample sizes.

Based on the discussion above, regardless of the level (national, regional and provincial) at which the model is formulated, the IWEE method generates the smallest estimated household level variance, while the ELL method generates the smallest estimated cluster level variance. We note that the estimated household level variance under the ELL method with heteroscedasticity model varies from one unit to another, hence, the mean value is reported, and that the estimated R^2 for including heteroscedasticity in the model is negligible, R^2=0.03 even at the national level, so that in terms of regression model fit at least it appears to offer few advantages for this data set.

Impact of the Variance Components on Prediction in SAE of Poverty

As elaborated in Section 2.3.3, in poverty estimation we are interested in area-level summaries of non-linear functions of $\hat{Y}_{\dot{b}h}$ in equation 2.13 rather than the regression fitting per se. It is instructive however to examine the effects of model uncertainty on area mean estimates

$$\bar{y}_a = \bar{\mathbf{X}}_a \hat{\boldsymbol{\beta}} \tag{2.21}$$

where $\bar{\mathbf{X}}_a$ is the population (or census) mean for area a of the covariates including the constant 1, after the regression model has been applied to the census data. By similarly averaging (2.3) to get the true mean $\bar{Y}_a$, subtracting from (2.21), and applying the variance operator, we get the prediction error variance equation:

$$V(\bar{y}_a - \bar{Y}_a) = \bar{\mathbf{X}}_a \boldsymbol{\Phi} \bar{\mathbf{X}}_a' + \frac{1}{N_a^2} \sum_{\dot{b}=1}^{\dot{B}} N_{\dot{b}}^2 \sigma_v^2 + \frac{1}{N_a} \sigma_e^2 \tag{2.22}$$

where N_a is the population size at a particular level of aggregation, $N_{\dot{b}}$ is the population size in each cluster, $\boldsymbol{\Phi}$ is the variance-covariance matrix of the regression coefficient estimates, and (σ_v^2, σ_e^2) are the cluster and household level variance components, respectively. Note that estimating this prediction error variance requires estimates of the variance components, but any bias caused by uncertainty in these would be a second order effect (see Prasad and Rao, 1990).

Based on (2.22), the extent of the influence of the survey based regression model and other variance components (cluster and household level) on the accuracy of the final small area estimates can be compared for any fitting technique and/or levels of aggregation. Generally, it is either the regression model (via the variance-covariance matrix of the estimated regression parameters) or the cluster effect that dominates the estimated precision of the computed small area estimate. Using the national level model in Table 2.1 and the survey data (instead of the census) auxiliary variables to generate (2.22), the results show that the extent to which the regression model effect contributes the most to the estimated variance of the prediction error increases markedly as household data are more aggregated - about 0.25% at the municipal level, 20% at the provincial level and 70% at the regional level. In other words, the more aggregated the data into larger areas, the greater the dominance of the regression model parameter uncertainty, regardless of the regression fitting method. This is as expected because even at high levels of aggregation, the contribution to the overall variance from the model effect depends on the average covariate values, not on the population size. This is the reason that, at the most aggregated regional level, small area techniques usually offer little improvement over direct estimates. This also justifies the necessity to examine in detail the regression fitting procedures applied in small area estimation of Third World poverty measures.

The effect of cluster level variation is different: at lower levels of aggregation (e.g., municipality) the computed variance of the prediction error of the small area estimates are dominated by the cluster component of variance or cluster level effect, i.e. for small areas (other than regional estimates) the variance component, not the regression model, has the greatest impact on the estimated variance of the prediction error of the small area estimates. Consequently, the accuracy of estimates of variance components especially at cluster level can be crucial to accurate estimation of standard error of small area estimates at the aggregation level at which they are most useful (for example at municipal level in the Philippines). From equation (2.22), it is easy to see that

$$\frac{\sigma_v^2}{\dot{B}} \leq \frac{1}{N_a^2}\sum_{\dot{b}=1}^{\dot{B}} N_{\dot{b}}^2\sigma_v^2 \leq \sigma_v^2 \tag{2.23}$$

so that small area estimates generally are more precise when the number of clusters $\dot{B}$ in each small area is larger.

Presented in Tables 2.4-2.6 are Kruskal-Wallis (KW) tests (Siegel, 1956) for the various fitting methods conducted on the estimated variances at the municipal (Table 2.4), provincial (Table 2.5) and regional (Tables 2.6) levels. In Table 2.4 significant differences exist among the variance estimates generated by the various small area techniques, as shown by the p-values of the Kruskal-Wallis statistics. Multiple comparison of mean ranks shows the Pseudo-EBLUP and IWEE methods have variance estimates at cluster level that are significantly higher than the other methods, but not significantly different from each other (although for the IWEE method the Z-value for the difference from average rank is in general rather higher than all the others).

The ELL and the GSR methods generate significantly lower and similar variance component estimates. This is principally because we used the ELL variance components estimation technique in generating variance components for the GSR method (because GSR does not usually estimate variance components), although the residuals we used were not identical for the two regression fitting methods. As expected, at the municipal level for which small area estimates were used in practice, the cluster effect (rather than regression coefficient uncertainty) is generally the dominant part of the small area variance estimates. Since the ELL and GSR methods have similar cluster level variance, their corresponding variance estimates at small area also tend to be similar. Explicitly, observe from Table 2.4 that the ranking of the variance estimates generally conforms with the ranking of the cluster effects.

In poverty estimation, estimates at higher levels of aggregation, such as those in Tables 2.5 and 2.6, are generally carried out for comparison with direct survey estimates at these more aggregated levels. Agreement between these results is often used to support those indicated for lower levels of aggregation. In addition, Tables 2.5 and 2.6 show that the estimated variances for the poverty estimates generated by the different techniques are not significantly different from each other at the provincial and regional levels, an effect that is partially due to the small number of provinces and even smaller number of regions. The variances and hence the standard errors may not be significantly different from each other, but it is worth noting that the GSR method tends to generate the smallest estimated variance for the regression model and in turn the smallest estimated variance of the prediction error of the small area estimates for poverty at the regional level, even though GSR generates higher variance for the individual regression coefficients' estimates (corresponding to the diagonal elements in the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$). As expected, at an even higher level of aggregation for all methods, the relative effect of the regression component is more pronounced.

2.4.5 Conclusion and Recommendations

As shown in the results given in the previous Section, regardless of which of the four methods are used, the regression parameter estimates were very similar. Differences in the estimated variance components however are noticeable. For the cluster level variance component, ELL gave the lowest estimate in general across the different levels of aggregation (national, regional and provincial) while the IWEE method gave the lowest estimate for the household level variance component. The more important issue then could be the possible underestimation of standard errors of parameter estimates and of variance components particularly at cluster level. We have shown that at the level (i.e. municipal) where small area estimation is actually used for aid allocation, the variance estimate of the small area tends to be dominated by the cluster level component of variance rather than by the sampling variability of the regression parameter estimates. It is then important that the cluster level component of variance be estimated properly. Since the ELL method generates the lowest estimated value for the cluster level variance component, consequently it generates a possibly underestimated variance of the small area estimates.

The GSR gave similar estimates of standard errors for the small area estimates to ELL despite having higher standard errors (and using a sound covariance matrix) for regression parameters. This is because the ELL method of variance component

Table 2.4: Kruskal-Wallis Test for Estimated Variances at the Municipal Level (N=1243)

SAE	Cluster Effect			Beta Effect			Variance		
Techniques	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.002843	2961.2(a)	-3.22	0.0002311	3067.3(ab)	-0.89	0.00318	2963.4(a)	-3.18
ELL(w/ hetero)	0.002843	2961.2(a)	-3.22	0.0002128	2802.0(c)	-6.72	0.00316	2930.8(a)	-3.89
Pseudo-EBLUP	0.003094	3229.4(b)	2.67	0.0002449	3257.5(ad)	3.28	0.00346	3241.3(b)	2.93
IWEE	0.003294	3426.9(b)	7.01	0.0002529	3364.5(d)	5.64	0.00366	3441.3(b)	7.32
GSR(Stata)	0.002843	2961.2(a)	-3.22	0.0002311	3048.7(b)	-1.3	0.00317	2963.1(a)	-3.18
Overall		3108			3108			3108	
KW Statistic	H=69.92	(P=0.000)		H=72.19	(P=0.000)		H=78.06	(P=0.000)	

Table 2.5: Kruskal-Wallis Test for Estimated Variances at the Provincial Level (N=83)

SAE	Cluster Effect			Beta Effect			Variance		
Techniques	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.0002518	200.3	-0.65	0.0001162	207.7	-0.03	0.00039	202.3	-0.48
ELL(w/ hetero)	0.0002518	200.3	-0.65	0.0001095	190.1	-1.52	0.00038	196.3	-0.99
Pseudo-EBLUP	0.000274	214.9	0.59	0.0001239	224.2	1.37	0.00042	217.1	0.78
IWEE	0.0002916	224.2	1.38	0.0001287	234.1	2.22	0.00045	227.8	1.68
GSR (Stata)	0.0002517	200.3	-0.65	0.00010	184	-2.04	0.00037	196.4	-0.98
Overall		208			208			208	
KW Statistic	H=2.82	(P=0.589)		H=10.61	(P=0.031)		H=4.48	(P=0.344)	

Table 2.6: Kruskal-Wallis Test for Estimated Variances at the Regional Level(N=16)

SAE	Cluster Effect			Beta Effect			Variance		
Techniques	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.000050	38.2	-0.45	0.000077	40.9	0.08	0.00013	39.3	-0.23
ELL(w/ hetero)	0.000050	38.2	-0.45	0.000073	35.1	-1.05	0.00012	37	-0.67
Pseudo-EBLUP	0.000055	42.6	0.4	0.000082	46.9	1.23	0.00014	44	0.67
IWEE	0.000058	45.3	0.93	0.000085	50.1	1.85	0.00015	46.6	1.17
GSR(Stata)	0.000050	38.2	-0.45	0.000070	29.6	-2.1	0.00013	35.6	-0.94
Overall		40.5			40.5			40.5	
KW Statistic	H=1.30	(P=0.861)		H=8.36	(P=0.079)		H=2.58	(P=0.630)	

estimation was adopted for GSR and as pointed out above, when there is less aggregation, the level at which most small area estimates are actually used, the cluster level variance component dominates.

The Pseudo-EBLUP and IWEE methods incorporate survey weights correctly (given a suitable choice of pseudo-likelihood) and gave larger (i.e., more conservative) estimates of cluster level variance components. This suggests that these two methods and particularly IWEE are among the best of the currently available methods, not necessarily for estimating regression equations (where availability of standard software may give GSR an advantage), but for estimating the crucial variance components.

Given the theoretical limitations of the survey fitting stage of the ELL method, it is recommended that the fitting procedure employed by the ELL method be replaced with the other methodologies considered - the Pseudo-EBLUP, IWEE, and the GSR method. These other methods have valid theoretical basis mathematically and the results generated can be clearly interpreted given the assumptions. The different methodologies when applied to complex weighted survey data from the Philippines indicate that for variance component estimation from survey data and hence for small area estimation at a fine level, Pseudo-EBLUP and particularly IWEE are likely to be better than the GSR or the ELL methods, although GSR is sound and easy to use because it is available in off-the-shelf software.

These important considerations however need to be predicated by adequate data cleaning, sound matching of possible regressor variables (in terms of mean, variance, and meaning) between survey and census data. Possible regression variables should also be properly selected and the limitations placed on subdividing survey data by small sample sizes should be appropriately identified, since all estimated standard errors for both regression parameter and small area estimates (regardless of estimation method and the model fitting used) are conditional on the regression model being correct.

2.5 Summary

This Chapter starts with a discussion of the poverty measures employed in Third World or developing countries and issues on the method of measuring poverty given the complexity and multi-dimensionality of the problem. This is followed by a discussion detailing the ELL method which is the most widely implemented small area estimation technique for poverty measures in developing countries. Theoretical limitations of the method were pointed out and its survey fitting method was compared with other existing methods such as the PEB, IWEE and GSR. The ELL and the other three methods were applied to the Philippine data. Results showed that the estimates of the regression parameters were very similar although differences were observed in the estimates of the variance components. In addition, the cluster level variance component was also observed to dominate the estimated variance of the small area estimates at the level of aggregation relevant for aid allocation.

Given the limitations of the ELL method, there are a number of ways in which the method could be improved. Two recommendations were made here: (1) an area level effect should be added to the existing income/consumption model and (2) the survey fitting stage of the method should be replaced with either the PEB, IWEE or GSR method.

In the framework of the ELL method for generation of small area estimates of poverty measures given the data available in Third World countries: a census (more detailed census information on auxiliary variables) and a survey (less detailed survey data which contains information on auxiliary variables and the variable of interest). It is assumed that the census and survey data have been conducted at the same time period, so that generation of small area poverty estimates becomes a problem when we have a new survey and a census conducted in an earlier period. The next Chapter gives a background of the different methods used for generating intercensal (period in between census or non-census years) small area estimates.

Chapter 3

Intercensal Updating of Small Area Estimates of Poverty Measures

3.1 Introduction

Updating population estimates for local areas is one of the primary concerns that led to the development of the earliest methods for small area estimation. A group of methods, so called "traditional demographic methods," is one set of methods developed in the 1950s. An example of a method under this group is presented in Section 3.2.1. Another method that emerged in the late 1970s and early 80s is the group of synthetic methods and is described in Section 3.2.2.

The more recent updating methods are also presented in this Chapter. These are small area updating techniques for income/expenditure-based poverty measures in Third World countries. With the need for more recent poverty estimates for policy making and aid allocation as well as to monitor poverty levels in order to assess progress towards the Millennium Development Goals, updating small area estimates of poverty is of utmost importance. In Section 3.3 updating techniques for small area estimation of poverty measures are presented including some applications, and a summary of the Chapter is presented in Section 3.4.

3.2 Traditional Updating Methods

3.2.1 Demographic Methods

Traditional demographic methods are the earliest techniques of small area estimation; they use demographic models (demographic variables are used as auxiliary variables, in conjunction with the latest census counts) to generate population estimates in local or small areas during intercensal years. One of the techniques under this category is the Symptomatic Accounting Techniques (SAT). It is called symptomatic in the sense that changes in demographic variables are strongly related to changes in local population (Rao, 2003). An example of the symptomatic method is the updating method proposed by Bogue (1950) also known as the *vital rates* (VR) method.

The VR method uses birth and death data as auxiliary variables for updating the estimate of the population for each small area (Y_{at}). It is assumed that the ratio of the birth (or death) rate for a given small area to the larger area's birth (or death) rate remains the same, $br_{at}/br_{pt} = br_{ao}/br_{po}$, where br_{at} and br_{pt} are birth rate for the current year in the small area and the larger area (where the small area is located), respectively, while br_{ao} and br_{po} are the corresponding birth rates for the last census. Moreover, the method assumes that the larger area birth rates and the number of births ($\tilde{b}_{at}$) for the local area a at time t are available from official sources. Hence, the estimate of the birth rate for the small area is: $\hat{br}_{at} = br_{pt}(br_{ao}/br_{po})$. The VR method estimate of the population total for small area a at the current year is:

$$\hat{Y}_{at} = \frac{\tilde{b}_{at}}{\hat{br}_{at}} \tag{3.1}$$

Whenever birth and death data are available the estimate of the current population of a small area is obtained as the average of the two small area estimates: $\hat{Y}_{at} = 1/2(\tilde{b}_{at}/\hat{br}_{at} + \tilde{d}_{at}/\hat{dr}_{at})$, where $\tilde{d}_{at}$ is the number of deaths in the small area and $\hat{dr}_{at}$ is the estimated death rate.

Marker (1999) and Noble (2003) showed that the VR model is an example of a linear model as follows:

$$Y_{at} = X_{ao}\beta + e_a \tag{3.2}$$

where $E(e_a)$=0 and $V(e_a) = \sigma^2 X_{ao}/w_a$, $Y_{at} = br_{at}$, $X_{ao} = br_{ao}$ and $w_a = N_{ao}/N_{po}$, and N_{ao} is the population in the small area at the last census while N_{po} is the population in the larger area at the last census. Model (3.2) is a special case of the GLM model presented in equation (1.2).

3.2.2 Indirect Procedures

Traditional indirect procedures are "techniques that use the values of the variable of interest from a domain and/or time period other than the domain and time period of interest" (Schaible, 1996). Estimators generated through traditional indirect procedures are generally design based and their design-variances are usually small relative to the the design variances of direct estimators. However, indirect estimators are generally design biased and the bias will not decrease as the overall sample size increases. The design bias will be small only if the implicit model is approximately true, leading to a smaller design mean square error (MSE) compared to the MSE of direct estimator. Reduction in MSE is the main reason for using indirect estimators (Rao, 2003).

One of the traditional indirect estimators is the *synthetic estimator*, this estimator can be generated if there is a reliable direct estimator for a larger area, covering several small areas under the assumption that the characteristics of the small areas are similar to the larger area (Gonzalez, 1973). If the said assumption is satisfied, then the synthetic estimator is generally very efficient - its MSE is small. However, it can be heavily biased for areas exhibiting strong individual effects which can lead to a larger MSE.

A synthetic small area estimator for small area a is given by

$$\hat{\bar{Y}}_a^{\mathsf{sy}} = \sum_{\tilde{g}} (N_{a\tilde{g}}/N_{a.})\bar{y}_{.\tilde{g}} \tag{3.3}$$

where $\bar{y}_{.\tilde{g}}$ is the reliable direct estimator of the larger domain ($\tilde{g}$) means, $N_{a\tilde{g}}$ are auxiliary information (which could be from the census) in the form of totals and $N_{a.} = \sum_{\tilde{g}} N_{a\tilde{g}}$. An example of this is the estimator proposed by Gonzalez and Hoza (1978) in generating employment estimates for counties in the United States both for the census year and intercensal years. Purcell and Linacre (1976) proposed a similar estimator, however, they used $N_{.\tilde{g}} = \sum_a N_{a\tilde{g}}$ for its denominator (by using this denominator, the synthetic estimator becomes a consistent estimator). The synthetic estimator can be viewed as an updating technique if the reliable direct estimator ($\bar{y}_{.\tilde{g}}$) is from the current survey and the auxiliary information comes from the census which is from a previous time period.

The synthetic estimator mentioned above can be expressed in terms of the linear model described in Chapter 1 as follows: $Y_{a\tilde{g}h} = X_{a\tilde{g}h}\beta_{\tilde{g}} + e_{a\tilde{g}h}$ where $Y_{a\tilde{g}h}$ denotes the hth unit in the small area a and subgroup $\tilde{g}$ (assumed to have available information from the sample), $X_{a\tilde{g}h}$ is an indicator variable for subgroup membership, $\beta_{\tilde{g}}$ is the mean value for subgroup $\tilde{g}$, and $e_{a\tilde{g}h}$ is an error term.

A generalization of synthetic estimation is the structure preserving estimation (SPREE) technique. It makes fuller use of direct estimates and uses the method of iterative proportional fitting (IPF) of margins (Purcell and Kish, 1980). IPF is used to adjust the cell counts of a contingency or multi-way table such that the adjusted counts satisfy specified margins. The cell counts are obtained from the last census while the specified margins represents reliable direct survey estimates of current margins. In this way, SPREE provides intercensal estimates of small area totals of characteristics also measured in the census (Rao, 2003).

As explained by Purcell and Kish (1980) there are two data sets required under the SPREE method: 1) the census data, which establishes the relationships between the variable of interest and the auxiliary variables at some previous time point, at the required small area level, and 2) the survey data, which establishes current relationships between the variable of interest and the auxiliary variables at the large domain level, accumulated over the small areas of interest. The objective of the estimation procedure is to conform to the current data in the survey data (updated margins) and to preserve the earlier relationships present in the census without interfering with the first objective. Extension of the SPREE methodology to a generalized linear model (GLM) based estimation procedure is proposed by Noble et al. (2002). This approach allows the SPREE method to be extended from the contingency table with categorical variables which the IPF could fit, to continuous variables and random effects model. Details of the SPREE method are presented in the next Chapter.

3.3 Updating Techniques for Small Area Estimates of Poverty Measures

There are two types of ELL-based updating technique that have recently been implemented. One is using panel survey data and the other one uses “time-invariant” variables. The panel data approach has been used in Uganda (Hoogeveen et al., 2003) and Thailand Jitsuchon and Lanjouw (2005) while the approach using time-invariant data has been implemented in Vietnam and Philippines (Lanjouw and van der Wiede, 2006). The panel data approach requires the availability of a longitudinal data set -

a data set such that data on individuals or households are gathered over time so that multiple observations are available on each individual or household in the sample. An example is the British Household Panel Survey (BHPS), carried out at the Institute for Social and Economic Research of the University of Essex (Taylor, 2001). The BHPS started in 1991 and the set of sampled households have since been followed and interviewed every year. The database from the BHPS is very popular and is usually used for research on social and economic change. The other approach on the other hand requires the availability of two cross-sectional surveys. Cross-sectional data refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time. Details and example of the two approach as applied to intercensal updating of small area poverty estimates are presented in the next two Sections.

3.3.1 Panel Data Approach in Updating Small Area Estimates

The panel data method requires collection of the most recent period (t_1) per capita income/expenditure for (a subset of) households included in the sample survey conducted in the same time period as the census (t_0). There are two techniques available for this methodology depending on the available covariates in the period (t_0) - household level or village level characteristics.

When household-level characteristics are available, updated welfare estimates are derived by combining the (t_1) per capita income/expenditure information y_{bh,t_1} with covariates or household characteristics that are common to the survey and the census collected in t_0 denoted by $\mathbf{x}_{bh,t_0}$ (Hoogeveen et al., 2003). The model used is as follows:

$$y_{bh,t_1} = \mathbf{x}_{bh,t_0}\boldsymbol{\beta} + u_{bh,t_1} \tag{3.4}$$

where $\boldsymbol{\beta}$ is the regression parameter and u_{bh,t_1} is the random error term. In this technique, new information on household characteristics are not needed to update poverty measures estimates, only the most recent information on income/expenditure is required as it uses the covariates from the census year.

In the implementation in Uganda, the two sets of survey data available for updating small area estimates of poverty measures cannot necessarily be considered as a panel data similar to the BHPS data described above. The two sets of data were basically two separate cross-sectional surveys. One set has been gathered at the same time period as the census which includes information on the household characteristics $(\mathbf{x}_{bh,t_0})$ and per capita income or expenditure (y_{bh,t_0}). The second set is the most recent survey which also contains $(y_{bh,t_1}, \mathbf{x}_{bh,t_1})$. However the two data sets contained some households that were identified as having taken part in the two successive surveys (matching households) and these were considered to form a panel data. In other words, the panel data were derived from two separate cross-sectional surveys that are not designed to generate panel data sets. One implication of the manner of implementation of the method is that there could be a substantial reduction in the amount of survey data (i.e. in the sample size) available which could result in less precise estimates. In Uganda, only about a thousand (1,071) households were part of the panel data out of the ten thousand households in the two surveys from which the panel data set was extracted.

In addition to the issue on the precision of the estimates generated, there are other limitations of this approach. In order for this method to generate valid and reliable estimates it assumes that 1) the most recent per capita income/expenditure (y_{bh,t_1}) depends on the household characteristics $\mathbf{x}_{bh,t_0}$ which are at least about 3 years before information on (y_{bh,t_1}) has been gathered and that 2) there is no net migration among small areas under consideration. Under this method, the estimates are generated by fitting the model given in equation (3.4) and following the ELL approach described in the previous Chapter.

In cases where household level panel data is not available, Jitsuchon and Lanjouw (2005) suggested the use of village level characteristics at the census period, $(\mathbf{x}_{a,t_0})$, as covariates. This approach is similar to the previous technique described, however, village-level data on explanatory variables from the census period are considered. A preliminary implementation of the updating approach in ten provinces of rural northeast Thailand used a single model relating the year 2002 per capita income (y_{bh,t_1}) to year 2000 village characteristics $(\mathbf{x}_{a,t_0})$. The model fitting approach and generation of updated estimates also follows the ELL method. The range of R^2 values computed for the fitted model (log per capita income) was 0.25 - 0.29, which is quite low, and is contradictory to expectation. The R^2 can be higher for the model fitted at higher level of aggregation of the dependent variable since household level errors are averaged out, however this can result in higher model error, and is difficult to implement for log transformed data such as log expenditure or log income.

3.3.2 Cross-sectional Surveys for Updating Small Area Estimates

In other Third World countries (e.g., Philippines) panel survey data is not always available; more commonly available are cross-sectional surveys conducted once every three years. Under this situation, Lanjouw and van der Wiede (2006) proposed an ELL-based method which employs the selection of "time-invariant" variables from the census period (t_0). This method has been used in a collaborative project of the World Bank and National Statistical Coordination Board (NSCB) on intercensal updating of small area estimates of poverty measures in the Philippines.

The implementation of this method is quite similar to the panel data technique (i.e., synthetic panel). Income/expenditure data taken from the most recent survey is combined with what are claimed to be time-invariant variables, common to the survey and census, collected in the census year to generate poverty measures estimates. The survey model is as follows,

$$y_{\dot{b}h,t_1} = \tilde{\mathbf{x}}_{\dot{b}h,t_0}\boldsymbol{\beta} + u_{\dot{b}h,t_1} \tag{3.5}$$

where $\tilde{\mathbf{x}}'_{\dot{b}h,t_0}$ refers to characteristics that are time invariant (i.e., $\tilde{\mathbf{x}}_{\dot{b}h,t_0} = \tilde{\mathbf{x}}_{\dot{b}h,t_1}$, in practice $\tilde{\mathbf{x}}_{\dot{b}h,t_1}$ is used for fitting the survey regression); $\boldsymbol{\beta}$ and $u_{\dot{b}h,t_1}$ are as defined in the previous Section.

Similar to the "panel data approach", the validity of this methodology also depends on two assumptions: 1) independent or explanatory variables in the survey model are time invariant, i.e., household characteristics and municipality/village means do not change from the census period to the most recent survey. The implementation in the Philippines so far did not consider any test for time invariance but only selected those variables that were deemed "logically" time invariant by the proponents of the method, e.g. at least high school educational attainment of household head.

Restricting to time-invariant variables however might result in a limited and hence poorly-fitting model. 2) migration (at least among small areas) between the census period and the most recent survey is negligible.

The implementation in the Philippines of the ELL-based updating method using time-invariant auxiliary variables is similar to the ELL method described in the previous Chapter with the addition of the two major assumptions mentioned above. As described in NSCB (2009), the auxiliary variables considered time invariant were:

- *household characteristics* - educational attainment of household head, type of family (extended family or not), and housing materials.

- *barangay characteristics* - barangay location (part of city/town proper or not); presence of various baranggay facilities (e.g., hospital, road networks, telephone system, electric power, among others); and barangay level statistics related to household characteristics (e.g., average family size, housing units floor area, and household members educational attainment) and business establishments (average number of hotels, dormitories, and other lodging places).

- *municipal characteristics* - municipal level statistics related to household characteristics (e.g., ownership of various home appliances, sanitary toilet, agricultural lands, and residential lands) and some other characteristics related to selected (e.g., at least 5 years old) individuals in the municipality (e.g., language spoken, employment, and school attendance).

The auxiliary variables used in fitting the ELL-based updating model are basically dominated by barangay and municipal level characteristics which are taken from the census data. Only the data on household characteristics are taken from the new survey. This could mean that the ELL-based updating method does not incorporate much updated information into the updating model. In addition, several ELL models for updating were fitted (see NSCB, 2009), one for each region (17 regions), which could lead to some problems of parsimony due to overfitting of models. Fitting several models for each region is the usual practice in implementing the ELL method as opposed to the use of one "global" model for the whole country in the implementation by Haslett and Jones (2005) in the Philippines. In Chapter 7 the updated estimates based on the ELL method using time-invariant variables are compared with the estimates generated from the proposed extended SPREE updating method described in Chapter 5.

3.4 Summary

Generation of updated estimates of local area population counts necessary for various government decisions is one of the important problems that led to various development not only in intercensal updating of population counts but in the small area estimation in general in various fields of application. The earliest methods for intercensal updating discussed in this Chapter are the demographic and synthetic methods. The group of demographic methods uses demographic variables as auxiliary variables for updating population estimates and in this Chapter the method using vital rates is described. The group of synthetic methods use survey data or survey-based or direct estimates of a larger area (containing the local or small area of interest) to generate the required small area statistics and is related to the SPREE method which is the basis of the proposed updating method in this thesis. Details of the SPREE method are discussed in the next Chapter.

Recent methods on updating small area estimates of poverty measures in Third World countries are also presented in this Chapter; these are all based on the ELL method described in Chapter 2. The ELL-based methods that have been implemented in some countries are either using panel data or time invariant variables. Comparison of the updated small area estimates of poverty measures in the Philippines based on the ELL method using time-invariant variables and the extended SPREE method proposed as an updating technique in this thesis is presented in Chapter 7. The description of the SPREE method and discussion of the method in the next Chapter is in the context of poverty estimation in Third World countries.

Chapter 4

The Structure Preserving Estimation Method

4.1 Introduction

The most recent methods proposed for intercensal updating of poverty measures in Third World countries discussed in Chapter 3 are all based on the framework of the ELL method. As described in Section 2.3.2, under the ELL method, a regression model is fitted to the survey data and the model is then applied to the census data to generate small area estimates. The ELL method however has issues in its theoretical underpinning as pointed out in Section 2.4.3, hence ELL-based updating method may not necessarily be reliable. An alternative updating method is therefore needed if more accurate updated estimates are to be generated. Our proposed updating method is an extension of the structure preserving estimation (SPREE) method. Hence, a detailed discussion of the SPREE method is presented in this Chapter.

SPREE was developed within the categorical data analysis framework by Purcell and Kish (1979) and applied to small area estimation. As demonstrated by Noble et al. (2002), SPREE models belong to the family of the *generalized linear models* (GLMs) described in Chapter 1. Description of the SPREE method which used the IPF algorithm is presented in Section 4.2. An example on poverty status as the variable of interest is presented for better understanding of the method. Explicit loglinear models for SPREE are presented in Section 4.3 including the model fitting procedure. The relationship between loglinear and logistic regression model is explored in Section 4.4 followed by the discussion of the modified SPREE in Section 4.5 which gives a prelude to the proposed intercensal updating method in the next Chapter. Section 4.6 discusses the generalized linear structural models (GLSM) proposed by Zhang and Chambers (2004) which is a generalization of the conventional SPREE model aimed to reduce the bias in SPREE.

4.2 The SPREE method and IPF

4.2.1 The SPREE method

SPREE is one of the categorical data analysis approaches to small area estimation. It implicitly fits loglinear models to a contingency table and uses the iterative proportional fitting (IPF) method for estimation. As described by Purcell (1979), there are two explicit assumptions on data availability that need to be satisfied for proper implementation of this methodology. First, current estimates (counts or relative frequencies) for the variable of interest (e.g., poverty status), cross-tabulated by appropriate associated variables (e.g., educational attainment, age etc.) should be available at the large domain level (i.e., aggregated over the small domains but possibly subdivided by other variables). This current data is generally available from large scale surveys like the surveys conducted by national statistical agencies in most countries. We note that these surveys may contain sample units from only a subset of the small areas. Second, estimates of the variable of interest cross-tabulated by the same associated variables should be available for each small domain at some previous time point. The small domain estimates are usually taken from the previous census or other reliable sources (Purcell, 1979). The SPREE method therefore requires two sets of data, the "census type" and the "survey type" data.

There are two important terms from Purcell (1979) that will be used in this Chapter and the Chapters that will follow - *association* and *allocation* structures. The structure inherent in the census type data is called association structure, this structure establishes the relationship between the variable of interest and the associated or auxiliary variables at some previous time point, at the required small area level. On the other hand, the structure inherent in the survey type data is called allocation structure; this structure establishes current relationships between the variable of interest and the auxiliary variables at the large domain level (aggregated in some way over the small areas). The two main objectives of the SPREE estimation procedure are:

(1) to update the census type data in order to conform to the information contained in the survey type or current data from sample survey in the allocation structure and

(2) to preserve the earlier relationships present in the association structure without interfering with the first objective.

The estimation process can be carried out using the iterative proportional fitting (IPF) algorithm developed by Deming and Stephan (1940). Details of the IPF method are presented in the next Section.

For a clearer understanding of the SPREE method, we present a simple example and introduce notations that will be used in this Chapter and in the next Chapters. We consider poverty status as our variable of interest, urbanity as the associated variable and provinces as the set of small areas of interest. The following notations will be used for convenience and clarity: Let a, b, and c denote the ath small area ($a = 1, ..., A$), bth variable of interest category ($b = 1, ..., B$), and cth associated variable category ($c = 1, ..., C$). We also let Y_{abc} be the required cell counts which are unknown; $\mathbf{p}$ be the set of relative frequencies of the required cell counts, such that its elements are: $p_{abc} = Y_{abc}/\sum_a \sum_b \sum_c Y_{abc}$; $\boldsymbol{\pi}$ is an $A \times B \times C$ array representing the association structure (relative cell frequencies established in the most recent census year) which for the standard version of SPREE is assumed to be fixed and not subject to error, and $\mathbf{p}^*$ a $B \times C$ matrix containing the allocation structure. We note that in this Chapter, census type or census data is used interchangeably with association structure and survey type data or survey margins with allocation structure.

Considering our example above, we assume that census type data is available cross-classified by poverty status (poor and non-poor), urbanity (urban and rural), and provinces. In addition, we assume that data on poverty status categorized by urbanity is available from a survey. If we put the information that we have in a diagram and for simplicity we assume that we only need to generate estimates for two small domains (e.g., two provinces), we will have the illustration in Figure 4.1 which is adopted from Purcell and Kish (1979).

The association structure is represented by eight cells labeled “Census”, while the allocation structure is represented by the margins labeled “Survey” contained in four cells - estimates are obtained on poverty status by urbanity cross-tabulation aggregated over the two provinces. Recent estimates at the province level are either not reliable or not available. To generate estimates of the variable of interest at the province

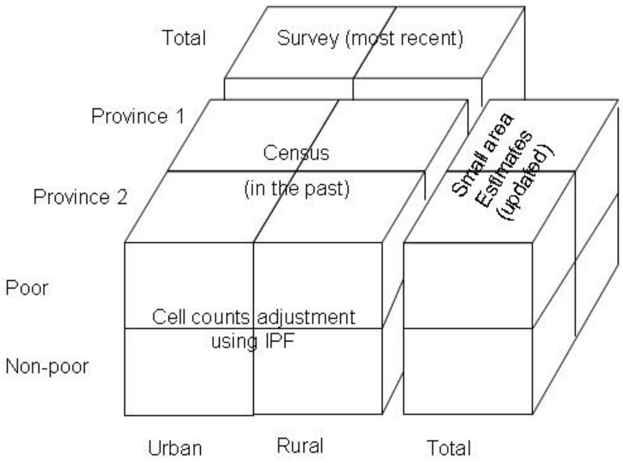


Figure 4.1: An example of data structure for SPREE method

level using SPREE, the association structure is adjusted (through IPF as will be illustrated in the next section) to the new margins while in some way preserving, as much as possible, the interaction structure between the variables as established in the census. The required estimates are then obtained by summing the adjusted table across the categories of urbanity. Based on the figure above, the required estimates are the four cells on the right hand margin, the poverty status by provinces summed over categories of urbanity.

We present in Figure 4.2 a more detailed illustration of the association structure with the necessary notations. We let I_{abci} be an indicator function from which we define a three dimensional contingency table of order $(A \times B \times C)$. The indicator function has a value 1 if unit i (i=1,...,N) is in domain a and has the response b for variable of interest and c for the associated variable, here N is the population size; and zero, otherwise. Letting $\{\tilde{Z}_{abc}\}$ be the census counts in a that have a response defined by b and c, then $\tilde{Z}_{abc} = \sum_{i=1}^{N} I_{abci}$. The observed relative frequency is then $\pi_{abc} = \tilde{Z}_{abc}/N$, which is shown in Figure 4.2.

For a simple illustration of the allocation structure $(\mathbf{p}^{*})$, we assume that we have new

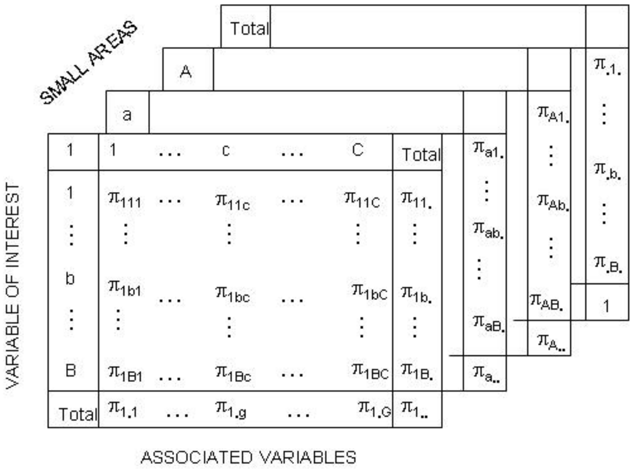
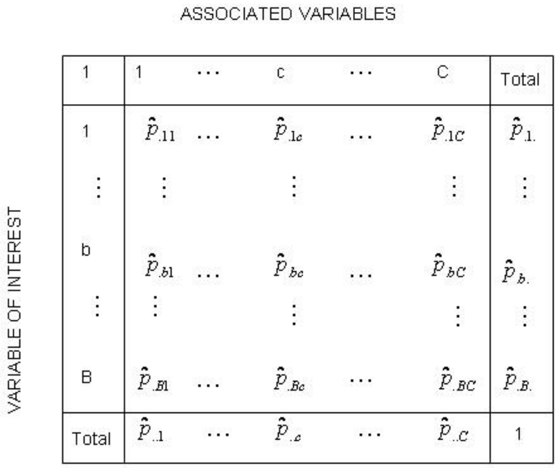


Figure 4.2: An example of data structure for SPREE method

information on the variable of interest and the associated variable. In this case, the allocation structure is simply the current structure inherent in the relative frequencies (accumulated in some way over the small domains) for the variable of interest, cross-tabulated by some associated variables. We let J_{bci} be an indicator function from which we define a two dimensional contingency table of order $(B \times C)$. The indicator function has a value 1 if the unit i has response b for the variable of interest and c for the associated variable and zero, otherwise. The relationship between the variable of interest and associated variable can then be summarized in a contingency table of relative frequencies $\hat{p}_{.bc} = \hat{Y}_{.bc} / \sum_b \sum_c \hat{Y}_{.bc}$ where $\hat{Y}_{.bc} = \sum_{i=1}^{N} W_i \varrho_i(s) J_{bci}$, W_i is the sampling weight (often the inverse of the selection probability) of unit i and $\varrho_i(s)$ is an indicator random variable characterizing the sampling design with value 1 if unit i is in the sample and zero otherwise. Here, the margins $Y_{.bc}$ are assumed to have sufficiently accurate estimates $(\hat{Y}_{.bc})$ from the survey. The allocation structure is shown in Figure 4.3.

Under SPREE methodology, the aim is to ensure that estimates are generated so that the constraint $\sum_a \hat{Y}_{abc} = \hat{Y}_{.bc}$ is satisfied, i.e., we require that the small area



ASSOCIATED VARIABLES

VARIABLE OF INTEREST

1	1	$\cdots$	c	$\cdots$	C	Total
1	$\hat{p}_{.11}$	$\cdots$	$\hat{p}_{.1c}$	$\cdots$	$\hat{p}_{.1C}$	$\hat{p}_{.1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
b	$\hat{p}_{.b1}$	$\cdots$	$\hat{p}_{.bc}$	$\cdots$	$\hat{p}_{.bC}$	$\hat{p}_{.b.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
B	$\hat{p}_{.B1}$	$\cdots$	$\hat{p}_{.Bc}$	$\cdots$	$\hat{p}_{.BC}$	$\hat{p}_{.B.}$
Total	$\hat{p}_{..1}$	$\cdots$	$\hat{p}_{..c}$	$\cdots$	$\hat{p}_{..C}$	1

Figure 4.3: An example of the allocation structure for SPREE method

estimates ($\hat{Y}_{abc}$) should sum to the known margins (or reliable estimates of margins) in the allocation structure. For the estimates generated under the case where the available allocation structure consisted of the bc margins, Purcell (1979) has shown that the weighted least squares, quasi-maximum likelihood and information-theoretic approaches all result in the same form of the estimator:

$$\hat{p}_{abc} = (\pi_{abc}/\pi_{.bc})\hat{p}_{.bc} \tag{4.1}$$

or $\hat{Y}_{abc} = (\tilde{Z}_{abc}/\tilde{Z}_{.bc})\hat{Y}_{.bc}$ which is similar to the synthetic estimator presented in Section 3.2.2. However, in cases where additional reliable information could be incorporated into the estimation process, a closed form of the estimator as above is not possible. The solution adopted by Purcell (1979) is the methodology called iterative proportional fitting (IPF) algorithm, a technique proposed by Deming and Stephan (1940) for a similar adjustment problem. Details of the IPF method are discussed in the next Section.

4.2.2 Iterative Proportional Fitting Algorithm

Deming and Stephan (1940) considered the problem of adjusting an estimated crosstabulation to agree with a number of known marginal totals, and their proposed method IPF is the first published instance of this method. IPF adjusts the marginal totals in an iterative cyclical fashion. Under SPREE for small area estimation, the IPF algorithm is used to adjust the census counts cross-classified by the small area, variable of interest and associated variables to agree with the marginal totals contained in the survey data or the allocation structure, thereby producing new estimates for each cell in the contingency table. The cell counts may be summed across the associated variables to get the required small area estimates. When there is only one set of constraints (in IPF available set of margins are considered constraints) in the allocation structure, a simple solution exists, and is a member of the general class of synthetic estimates presented in Chapter 3.

For illustration purposes, we assume that aside from the two-way allocation structure described in Figure 4.3, we also have current reliable information on the small areas. Using the notations of Purcell (1979), the IPF algorithm is implemented as follows:

(1) The starting values are set equal to the known past or census values,

$$\hat{p}^{(0)}_{abc} = \pi_{abc}$$

(2) The cell proportions are then adjusted to the first set of marginal constraints, specified by the allocation structure $\sum_h \hat{p}_{abc} = \hat{p}_{.bc}$, i.e.,

$${}_1\hat{p}^{(1)}_{abc} = \frac{\hat{p}^{(0)}_{abc}}{\hat{p}^{(0)}_{.bc}} \hat{p}_{.bc}$$

(3) The adjusted cell values from the previous step are then adjusted to the second set of marginal constraints, $\sum \sum_{bc} \hat{p}_{abc} = \hat{p}_{a..}$, i.e.,

$$\hat{p}^{(1)}_{abc} = \frac{{}_1\hat{p}^{(1)}_{abc}}{{}_1\hat{p}^{(1)}_{a..}} \hat{p}_{a..}$$

Steps (2) and (3) are then repeated in a cyclical fashion. In general, at each kth iteration, the resulting estimates are $\hat{p}^{(k)}_{abc}$ are used as inputs into the next cycle. The iteration is continued until some convergence criterion is satisfied. This method can be extended to any finite number of dimensions and as pointed out by Bishop et al. (1975) convergence is guaranteed if the margins are consistent.

4.3 Loglinear Models and the SPREE method

4.3.1 Loglinear Models for SPREE

The SPREE-based small area estimation, as mentioned in the previous Section, is in the framework of categorical data analysis that involves two multi-way contingency tables (one for the census or auxiliary variables and one for the survey data) and the relationships among the three sets of variables - variable of interest, small area and associated variables. For the SPREE method via explicit fitting of loglinear models, two loglinear models are involved - one for each of the contingency tables. To illustrate the loglinear models, we will use the definition of variables presented in the previous Section. As defined earlier, we let Y_{abc} denote the set of required cell counts in the contingency table corresponding to the small area a, variable of interest category b, and associated variable category c in the survey, with $E(Y_{abc}) = \mu^{Y}_{abc}$. The set of counts $\{Y_{abc}\}$ are unknown, however some of the marginal totals, such as $Y_{.bc}$, are known or reliable estimates are available from a survey. The corresponding cell counts available in the census or administrative data are denoted by $\tilde{Z}_{abc}$ with $E(\tilde{Z}_{abc}) = \mu^{\tilde{Z}}_{abc}$. Loglinear model formulas generally use the counts, e.g., μ^{Y}_{abc}, hence Poisson sampling for the cell counts is assumed. The loglinear model corresponding to the three-way contingency table for the survey is as follows:

$$\log(\mu^{Y}_{abc}) = \lambda^{s}_{1} + \lambda^{s}_{a} + \lambda^{s}_{b} + \lambda^{s}_{c} + \lambda^{s}_{ab} + \lambda^{s}_{ac} + \lambda^{s}_{bc} + \lambda^{s}_{abc} \qquad (4.2)$$

where:

$$\begin{array}{rcl}
\lambda_1^s & = & (ABC)^{-1}\sum_{a,b,c}\log(Y_{abc}) \\
\lambda_a^s & = & (BC)^{-1}\sum_{b,c}\log(Y_{abc}) - \lambda_1^s \\
\lambda_b^s & = & (AC)^{-1}\sum_{a,c}\log(Y_{abc}) - \lambda_1^s \\
\lambda_c^s & = & (AB)^{-1}\sum_{a,b}\log(Y_{abc}) - \lambda_1^s \\
\lambda_{ab}^s & = & (C)^{-1}\sum_{c}\log(Y_{abc}) - \lambda_a^s - \lambda_b^s - \lambda_1^s \\
\lambda_{ac}^s & = & (B)^{-1}\sum_{b}\log(Y_{abc}) - \lambda_a^s - \lambda_c^s - \lambda_1^s \\
\lambda_{bc}^s & = & (A)^{-1}\sum_{a}\log(Y_{abc}) - \lambda_b^s - \lambda_c^s - \lambda_1^s \\
\lambda_{abc}^s & = & \sum_{a}\log(Y_{abc}) - \lambda_a^s - \lambda_b^s - \lambda_c^s - \lambda_{ab}^s - \lambda_{bc}^s - \lambda_{ac}^s - \lambda_1^s
\end{array}$$

For the census, on the other hand, the loglinear model is:

$$\log(\mu_{abc}^{\tilde{Z}}) = \lambda_1^o + \lambda_a^o + \lambda_b^o + \lambda_c^o + \lambda_{ab}^o + \lambda_{ac}^o + \lambda_{bc}^o + \lambda_{abc}^o \tag{4.3}$$

where the λ^o terms are defined similarly as the λ^s and all the λ-terms satisfy the constraints $\sum_a \lambda_a = \sum_b \lambda_b = \sum_c \lambda_c = \sum_a \lambda_{ab} = \sum_b \lambda_{ab} = \sum_a \lambda_{ac} = \sum_c \lambda_{ac} = \sum_b \lambda_{bc} = \ldots = \sum_c \lambda_{abc} = 0$. If we only have reliable estimates of the survey margins for the variable of interest and the associated variable (i.e., $\{\hat{Y}_{.bc}\}$, the SPREE estimation method is equivalent to fitting the loglinear model (4.3) then re-estimating the set of parameters $\{\lambda_1^o, \lambda_b^o, \lambda_c^o, \lambda_{bc}^o\}$ using the information available from the survey margins $\{\hat{Y}_{.bc}\}$. This process is also equivalent to fitting the loglinear model (4.2) and then setting those parameters that cannot be estimated (due to insufficient information) from the survey equal to the values generated from the census model (4.3), i.e, $\lambda_a^s = \lambda_a^o$, $\lambda_{ab}^s = \lambda_{ab}^o$, $\lambda_{ac}^s = \lambda_{ac}^o$, and $\lambda_{abc}^s = \lambda_{abc}^o$.

Expressing the loglinear model in (4.2) in the form of the generalized linear model discussed in Chapter 1, we let $\mathbf{Y}$ be the matrix of cell counts in a contingency table and $\boldsymbol{\mu}^Y$ be the corresponding expected value, the loglinear model in (4.2) is equivalent to:

$$\log(\boldsymbol{\mu}^Y) = \mathbf{X}\boldsymbol{\beta}^{(s)} \tag{4.4}$$

where $\boldsymbol{\beta}^{(s)}$ is a $(\dot{p} \times 1)$ column vector of parameters with $\dot{p}$ being the number of parameters in the model, and $\mathbf{X}$ is the $(\tilde{N} \times \dot{p})$ design matrix, where $\tilde{N}$ is the number of cells. The model under consideration is a saturated loglinear model hence, we will have $\tilde{N} = \dot{p}$. As illustrated by Noble et al. (2002), the design matrix and the vector of parameters are then partitioned into two parts as follows:

$$\log(\boldsymbol{\mu}^Y) = [\mathbf{X}_1 : \mathbf{X}_2][\boldsymbol{\beta}_1^{(s)} : \boldsymbol{\beta}_2^{(s)}]' = \mathbf{X}_1\boldsymbol{\beta}_1^{(s)} + \mathbf{X}_2\boldsymbol{\beta}_2^{(s)} \tag{4.5}$$

The second term in the rightmost side of equation above, $\mathbf{X}_2\boldsymbol{\beta}_2^{(s)}$, corresponds to that part of the design matrix and set of parameters that cannot be estimated due to insufficient information from the survey data. A similar model can be specified for (4.3) as follows,

$$\log(\boldsymbol{\mu}^{\tilde{Z}}) = \mathbf{X}\boldsymbol{\beta}^{(o)} \tag{4.6}$$

this equation can also be expressed as

$$\log(\boldsymbol{\mu}^{\tilde{Z}}) = [\mathbf{X}_1 : \mathbf{X}_2][\boldsymbol{\beta}_1^{(o)} : \boldsymbol{\beta}_2^{(o)}]' = \mathbf{X}_1\boldsymbol{\beta}_1^{(o)} + \mathbf{X}_2\boldsymbol{\beta}_2^{(o)} \tag{4.7}$$

The term $\mathbf{X}_1\boldsymbol{\beta}_1^{(o)}$ corresponds to that part of the design matrix and set of parameters for which information is available from the survey data and hence, in fitting the model, the said parameters will be re-estimated. The term $\mathbf{X}_2\boldsymbol{\beta}_2^{(o)}$, corresponds to that part of the design matrix and parameters that will remain unchanged, i.e., $\mathbf{X}_2\boldsymbol{\beta}_2^{(s)} = \mathbf{X}_2\boldsymbol{\beta}_2^{(o)} = \mathbf{X}_2\boldsymbol{\beta}_2$. It follows that equation (4.5) is then equivalent to

$$\log(\boldsymbol{\mu}^Y) = \mathbf{X}_1\boldsymbol{\beta}_1^{(s)} + \mathbf{X}_2\boldsymbol{\beta}_2 \tag{4.8}$$

Here, $\mathbf{X}_1\boldsymbol{\beta}_1^{(s)}$ corresponds to $\{\lambda_1^s, \lambda_b^s, \lambda_c^s, \lambda_{bc}^s\}$, while $\mathbf{X}_2\boldsymbol{\beta}_2$ corresponds to $\{\lambda_a^s, \lambda_{ab}^s, \lambda_{ac}^s, \lambda_{abc}^s\}$ which by assumption are set equal to $\{\lambda_a^o, \lambda_{ab}^o, \lambda_{ac}^o, \lambda_{abc}^o\}$. In terms of the GLM framework, the SPREE method is equivalent to fitting model (4.8).

4.3.2 Model fitting for Loglinear Models

To illustrate the model fitting procedure, we will consider the GLM given in (4.4) from the previous Section. To simplify the derivation, we will assume that the sampling model is Poisson and we let $\mathbf{Y}$ be the vector of observed counts such that $\mathbf{Y} = (y_1, \ldots, y_{\tilde{N}})'$ with $y_{\dot{a}}$ representing the $\dot{a}$th cell count such that $\dot{a} = 1, \ldots, \tilde{N}$ and $\tilde{N}$ is the total number of cells in a contingency table, note that the dot in $\dot{a}$ is used so that this notation will not be confused with a that is used to denote small areas in other Chapters. We also have $\boldsymbol{\mu}^Y = E(\mathbf{Y})$ such that $\boldsymbol{\mu}^Y = (\mu_1^Y, \ldots, \mu_{\tilde{N}}^Y)$. Model (4.4) can then be expressed as $\log(\mu_{\dot{a}}^Y) = \sum_{\dot{c}} x_{\dot{a}\dot{c}}\beta_{\dot{c}}$ for all $\dot{a}$, where $\dot{c} = 1, \ldots, \dot{p}$. For Poisson sampling the log likelihood is

$$\begin{aligned} L(\boldsymbol{\mu}^Y) &= \sum_{\dot{a}} y_{\dot{a}} \text{log} \mu_{\dot{a}}^Y - \sum_{\dot{a}} \mu_{\dot{a}}^Y \\ &= \sum_{\dot{a}} y_{\dot{a}} (\sum_{\dot{c}} x_{\dot{a}\dot{c}} \beta_{\dot{c}}) - \sum_{\dot{a}} exp(\sum_{\dot{c}} x_{\dot{a}\dot{c}} \beta_{\dot{c}}) \end{aligned}$$

The sufficient statistic for $\beta_{\dot{c}}$ is its coefficient, $\sum_{\dot{a}} y_{\dot{a}} x_{\dot{a}\dot{c}}$. The partial derivative of the log likelihood with respect to $\beta_{\dot{c}}$ is as follows:

$$\frac{\partial L(\boldsymbol{\mu}^Y)}{\partial \beta_{\dot{c}}} = \sum_{\dot{a}} y_{\dot{a}} x_{\dot{a}\dot{c}} - \sum_{\dot{a}} \mu_{\dot{a}}^Y x_{\dot{a}\dot{c}}$$

since

$$\frac{\partial}{\partial \beta_{\dot{c}}} [exp(\sum_{\dot{c}} x_{\dot{a}\dot{c}} \beta_{\dot{c}})] = x_{\dot{a}\dot{c}} exp(\sum_{\dot{c}} x_{\dot{a}\dot{c}} \beta_{\dot{c}}) = x_{\dot{a}\dot{c}} \mu_{\dot{a}}^Y$$

Equating the derivative of the likelihood equations to zero it will be of the form $\mathbf{X'Y} = \mathbf{X'}\hat{\boldsymbol{\mu}}^Y$, which equates the sufficient statistics to the corresponding expected values. If we consider a two way contingency table with observed counts $y_{\dot{a}}$ and a saturated loglinear model is fitted, the solution to the likelihood equation gives $\hat{\mu}_{\dot{a}}^Y = y_{\dot{a}}$. However, not all loglinear models have direct estimates or explicit formulas for maximum likelihood estimates; in these cases the maximum likelihood estimation process then requires iterative methods such as the Newton-Raphson algorithm (see Agresti, 2002) to solve the likelihood equations or the iterative proportional fitting algorithm described in Section 4.2.2. One of the differences between the Newton-Raphson and IPF method is that the IPF method does not generate the model parameter estimates and the estimated covariance matrix directly, it generates the fitted values and by using the said values generates the required model information.

The estimation procedure for the loglinear model parameters discussed above is not directly applicable when we are dealing with data from a complex survey. One of the methods suggested by Lohr (1999) to incorporate the sampling design is to generate the estimates of model parameters by using the sampling weights, i.e., generate estimates of cell proportions using the sampling weights and generate estimates of model parameters using the weighted cell proportions. The estimated variance of the model parameters is generated by any of the replication or random groups methods, i.e., refitting the loglinear model on each of the replicates and the variability among the parameter estimates from different replicates is used as the estimated variance (design-based) of the model parameter estimates. Similar procedure can be applied to generate the variance of the estimated counts from the fitted loglinear model.

Under standard SPREE the census data is assumed fixed. Hence, the only source of variation for the small area estimates generated through the SPREE method is the uncertainty in survey margins. The variance of the small area estimates generated via SPREE is the design-based variance mentioned above, that is generated through random groups or replication methods. More details of the variance estimation methods that can be used for SPREE is discussed in Chapter 6.

4.3.3 Model fitting for SPREE

Fitting loglinear models for SPREE for generation of small area estimates can be cumbersome and tedious as it entails fitting two loglinear models, one for the survey data and one for the census data, and this is especially cumbersome when dealing with many explanatory variables and large data sets (e.g., national survey and census). One method that could be used is called *table standardization* described by Agresti (2002) which is equivalent to the IPF method, in which, for counts Y_{ab} with $\mu^Y_{ab} = E(Y_{ab})$, the standardization process corresponds to fitting the model:

$$\log(\frac{\mu^Y_{ab}}{\mu^{\tilde{Z}}_{ab}}) = \mathbf{X}_1 \Delta \boldsymbol{\beta}_1$$

which is equivalent to fitting to the survey data the following model,

$$\log \mu^Y_{ab} = \mathbf{X}_1 \Delta \boldsymbol{\beta}_1 + \log \mu^{\tilde{Z}}_{ab} \tag{4.9}$$

where $\mu^{\tilde{Z}}_{ab}$ are the expected counts in the census contingency table and $\Delta\boldsymbol{\beta}_1 = \boldsymbol{\beta}^{(s)}_1 - \boldsymbol{\beta}^{(o)}_1$.

Fitting model (4.9) is equivalent to fitting a loglinear model to the survey data with $\log \mu^{\tilde{Z}}_{ab}$ as an *offset*. An offset is a predictor variable with a coefficient set equal to one. Setting the log of the counts from the census data, $\tilde{Z}_{ab}$, as an offset, is no different from setting the log of the predicted values from a saturated census model as an offset

(Qiao, 2006). Moreover, setting the logarithms of the predicted values of the census model as an offset is equivalent to defining $\mathbf{X}_2\boldsymbol{\beta}_2$ in model (4.8) as an offset. Hence, fitting model (4.9) is equivalent to fitting (4.8). That is,

$$\begin{aligned}\log\mu_{ab}^{Y} &= \mathbf{X}_1\Delta\boldsymbol{\beta}_1 + \log\mu_{ab}^{\tilde{Z}}\\ &= \mathbf{X}_1(\boldsymbol{\beta}_1^{(s)} - \boldsymbol{\beta}_1^{(o)}) + (\mathbf{X}_1\boldsymbol{\beta}_1^{(o)} + \mathbf{X}_2\boldsymbol{\beta}_2^{(o)})\\ &= \mathbf{X}_1\boldsymbol{\beta}_1^{s} + \mathbf{X}_2\boldsymbol{\beta}_2\end{aligned}$$

This fitting technique requires the use of pseudo-values because the cell level data is not available in the survey. These pseudo-values which we will refer to here as *pseudo-counts* ($\hat{\hat{Y}}_{ab}$) are computed from the available survey margins . These values are generated based on the structure of the design matrix $\boldsymbol{X}_1$. For example, if we have two sets of reliable margins $y_{a.}$ and $y_{.b}$ for two variables such that one has A categories and the other has B categories, the pseudo-counts will be

$$\hat{\hat{Y}}_{ab} = \frac{y_{a.}y_{.b}}{y_{..}}$$

As defined in the other Sections of this Chapter, $a = 1, \ldots, A$ denotes areas, $b = 1, \ldots, B$ denotes categories of the variable of interest and $(\cdot)$ represent a sum over that index. The computed pseudo-counts do not necessarily satisfy the assumption of independence; these values depend on the combination of survey margins available for estimation. Examples of computing pseudo-counts corresponding to specific loglinear models are presented in Section 8.6 of Agresti (2002).

4.4 The Logistic Regression Model for SPREE

There are cases when one of the variables in the loglinear model has only two levels, in particular poor and non-poor. This model can be expressed as a binomial (or binary) logistic regression which is a form of regression used when the response variable, Y is dichotomous and the explanatory variables $\boldsymbol{X}$ are of any type. Thus, this approach models directly the prevalence or incidence of poverty. Let $p(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$, then the logistic regression model is

$$p(x) = \frac{exp(\alpha + \beta(x))}{1 + exp(\alpha + \beta x)}$$

Equivalently, the log odds or *logit*, has the linear relationship

$$\text{logit}[p(x)] = \log\left[\frac{p(x)}{1-p(x)}\right] = \alpha + \beta x$$

This equates the logit link function to the linear predictor.

The model above describes how a single categorical response variable depends on a set of explanatory variables. A loglinear model on the other hand, treats categorical response variables symmetrically and describes associations and interactions among the variables. The two models seem different, but connections exist between them. For a loglinear model, forming logits on one response helps interpret the model. Moreover, logit models with categorical explanatory variables have equivalent loglinear models as will be shown in an example below.

The loglinear models in (4.2) and (4.3) have a general form

$$\log(\mu_{abc}) = \lambda + \lambda_a + \lambda_b + \lambda_c + \lambda_{ab} + \lambda_{ac} + \lambda_{bc} + \lambda_{abc}$$

This loglinear model is equivalent to defining a logit on one of the response variables, say poverty status (B) which has two categories (poor= 1 and non-poor= 0), with the following form:

$$\text{logit}[P(B = 1\backslash A = a, C = c)] = \alpha + \beta_a + \beta_c + \beta_{ac}$$

There is a direct relationship between the loglinear parameters and logit parameters as will be illustrated below. The expression above can be written as,

$$\begin{aligned}
\text{logit}[P(B = 1\backslash A = a, C = c)] &= \log\left[\frac{P(B = 1\backslash A = a, C = c)}{P(B = 0\backslash A = a, C = c)}\right] \\
&= \log[P(B = 1\backslash A = a, C = c)] - \log[P(B = 0\backslash A = a, C = c)] \\
&= \log(\mu_{a1c}) - \log(\mu_{a0c}) \\
&= (\lambda_{b=1} - \lambda_{b=0}) + (\lambda_{a,b=1} - \lambda_{a,b=0}) \\
&\quad + (\lambda_{b=1,c} - \lambda_{b=0,c}) + (\lambda_{a,b=1,c} - \lambda_{a,b=0,c}) \\
&= \alpha + \beta_a + \beta_c + \beta_{ac}
\end{aligned}$$

Hence, we can also use logistic regression models in lieu of the loglinear models for generating small area estimates under the SPREE framework at least for a binary response variable. The logistic regression model also helps in the interpretation of what is borrowed from the census model. For example if we have available survey margins for B (poor and non-poor) and C (urbanity) as well as BC (poor and non-poor cross-classified by urbanity) then in the equation above α and β_c are updated but β_a and β_{ac} are not, which are then borrowed from the census or equated to their corresponding census values. This means that the main effect of urbanity on poverty is re-estimated, but the interaction with area is not.

4.5 Modified SPREE

As described above, the SPREE method requires that the variable of interest be measured or available in both the survey and the census. However, there are instances when there is no information on the variable of interest in the census and this happens for example in the Philippines, where information on poverty status of households is not collected in the census. In these cases, the full association structure, as originally defined in Section 4.2.1 is not available for the estimation process. Purcell (1979) suggested an approach which artificially constructs a full association structure, i.e., a dummy association structure and then utilizes the usual estimation procedure for SPREE. This estimation procedure leads to the standard synthetic estimate by Purcell and Linacre (1976),

$$\hat{p}_{ab} = \sum_{c} \frac{\pi_{a.c}}{\pi_{..c}} \hat{p}_{.bc}$$

which is equivalent to $\hat{Y}_{ab} = \sum_c (Z_{a.c}/Z_{..c})\hat{Y}_{.bc}$ and is similar to the synthetic estimator presented in Chapter 3. This estimate implicitly assumes that there is no interaction between the variable of interest and the small domains, at each level of the associated variable, and that there is no three-way interaction between the variables. These assumptions are much more stringent than the implicit assumptions resulting out of adjustment with respect to the full association structure. Consequently, the biases of the estimates based on the dummy association structure are expected to be greater due to likelihood that these stringent assumptions may not be adequately met i.e., there are no higher order effects in the model.

The approach described here for estimation when the association structure is incomplete is just one of the possible alternative methods. In the next Chapter, an approach to deal with a similar problem is presented which is also our proposed intercensal updating method, an extension of the SPREE approach.

4.6 The Generalized Linear Structural Models

In Section 4.3.1 SPREE was described as fitting a loglinear model (4.8) to the survey data cross-classifications with all the terms $\mathbf{X}_2\boldsymbol{\beta}_2$ assumed equal to the corresponding terms in the model fitted to the census cross-classifications. A generalization of the loglinear model underpinning SPREE is proposed by Zhang and Chambers (2004) and is called the generalized linear structural model (GLSM). Under the GLSM, the assumption of equality between $\mathbf{X}_2\boldsymbol{\beta}_2^{(s)}$ and $\mathbf{X}_2\boldsymbol{\beta}_2^{(o)}$ is relaxed through a parameter called the proportionality coefficient (ζ). The parameter ζ essentially re-scales all the terms in $\mathbf{X}_2\boldsymbol{\beta}_2^{(o)}$, i.e., $\mathbf{X}_2\boldsymbol{\beta}_2^{(s)} = \zeta\mathbf{X}_2\boldsymbol{\beta}_2^{(o)}$. SPREE is then considered as a special case of GLSM such that $\zeta = 1$. An extended version of the model called generalized linear structural mixed model (GLSMM) is also proposed to allow for variability in the groups of parameters by adding a random error component to the model. Saei et al. (2005) proposed new models analogous to the GLSM and GLSMM by assuming a product-multinomial sampling distribution for the survey data.

The parameter ζ accounts in part for the changes in the association structure, leading to a reduction in bias. The introduction of the new parameter also affects the variance of the estimates that will be generated from the model, i.e. variance associated with $\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$. The variance would be scaled down if the estimate of ζ is less than one, and would be scaled up if $\zeta > 1$. This proportionality coefficient is assumed constant for all the parameters in the cell level log linear model for the census cross-classification that are not re-estimated using the survey data. Fitting the GLSM or GLSMM is relatively complicated but the underlying principle is similar to fitting a model to cell level parameter (i.e. loglinear parameter) estimates from the survey data with the corresponding cell level parameter estimates from the census as auxiliary variables in a simple linear regression framework.

Considering the method of fitting the GLSM or the procedure to estimate ζ, it can be deduced that the estimate of ζ can be highly influenced by some groups of parameters. For example, in the three way table - 90x3x2 used by Zhang and Chambers (2004) to illustrate the theory of GLSM the numbers of second and third order parameters of the loglinear model fitted for the table are 269 and 178 respectively. The estimate of the parameter ζ therefore involves a mix of second and third order parameters. For multiway tables of higher dimension, the number of parameters at each level in higher order tables yields even higher percentages of the higher level interactions. Hence this type of model is less attractive unless a set of ζ values could be fitted, e.g. one at each level in the hierarchy.

In addition, the GLSM model have some issues related to parameter estimability which is also described in detail by Haslett et al. (2007). For higher dimensional tables it is difficult to obtain parameter estimates and thus to scale by ζ, even when it can be estimated. Moreover, estimation can be very complicated in the presence of cells with zero frequency in both the survey and the census cross-classification data. Having cells with zero frequency would result in log odds ratio with zero in either the numerator or denominator or both. This situation could get worse with higher order parameters.

The problem of cells with zero frequency could limit the applicability of the GLSM to Third World small area updating of poverty measures which generally uses a relatively sparse survey data, and hence, would generally have zero frequency cells. For example in the Philippines, the sample size of the 2003 survey data is 42094 households and one of the possible cross-classification that can be used for updating is a 1623 municipalities $\times$ 2 poverty status $\times$ 4 wall type $\times$ 4 roof type $\times$ 2 urbanity $\times$ 2 sex of household head $\times$ 2 high school education attainment $\times$ 2 presence of domestic helper which results to a total of 830976, so that the average frequency or count per cell is around 0.05.

Aside from cells with zero frequency that can cause intractable parameter inestimability, some cells will have very small estimated values that could contribute to less stable loglinear parameter estimates. And less stable estimates especially for higher order effects are the ones that would highly influence the estimate of the proportionality coefficient. This again supports the argument on the simplicity of the assumption of a constant value for the proportionality coefficient.

Moreover, modeling under GLSM involves fitting a loglinear model to the census data that could require considerably more computing time than the classical SPREE and its extensions. Under the classical SPREE or its extensions, an offset can be used and do not provide parameter estimates directly, only cell estimates, hence shorten the computing time.

Overall, the GLSM is a very useful generalization of SPREE and has the great potential for reducing bias in the SPREE model for small area estimation. However, estimation of the parameter ζ can be very cumbersome if not impossible computationally for large data sets with many variables and since ζ is an average it may not significantly reduce bias in such circumstances. Moreover, applicability of the GLSM model is limited to using fewer explanatory variables in the model, which may not generally be useful for poverty estimation in most Third World countries. Given the limitations of the GLSM for large multiway tables, other methods of extending the classical SPREE method may be more applicable for small area estimation of Third World poverty measures.

4.7 Summary

This Chapter provides a detailed review of the SPREE method including the IPF method of generating SPREE-based small area estimates. It is emphasized that SPREE-based small area estimates generated through IPF can also be generated by fitting a loglinear model explicitly to the association and allocation structure. Implementation of the method involves fitting the loglinear model to the survey data (allocation structure) and then having the census (association structure) as an offset. The relationship between loglinear and logistic regression model is also described for cases when one of the auxiliary variables used in the loglinear model has only two categories. In such cases, either a loglinear or logistic regression model can be used for generating small area estimates under the SPREE framework.

The modified SPREE proposed by Purcell (1979) is also presented. This method is used to generate SPREE-based small area estimates when the variable of interest is not measured in the census, which is the case for the poverty estimation problem in the Philippines as well as in updating small area poverty estimates. As will be discussed in the next Chapter the method proposed uses a pseudo-census data instead of a dummy association structure as in the modified SPREE.

The generalized linear structural models proposed by Zhang and Chambers (2004) are also discussed. These models account for the bias in SPREE-based small area estimates generated under the assumption of a fixed association structure (i.e. association structure is assumed constant in the census period). A proportionality coefficient estimated from the data is introduced to account for the changes in the association structure leading to a reduction in bias. An alternative to GLSM, which is the method proposed for small area updating, and is also an extension of the SPREE method, is presented in the next Chapter. This method can also reduce the bias by allowing for a richer, higher dimensional table and hence model to be fitted, i.e. more explanatory variables are incorporated into the model, and also by allowing the association structure to be stochastic. The stochastic association structure is established by assuming that the census data is drawn from a superpopulation.

Chapter 5

Extended SPREE for Updating Small Area Estimates

5.1 Introduction

Background on small area updating methods for poverty measures in Third World countries was discussed in Chapter 3. As pointed out, those updating methods have various limitations in terms of data requirements and are based on the ELL method which also has issues in its theoretical underpinning as discussed in Chapter 2. In this Chapter, we develop an updating method which is an extension of the SPREE method described in Chapter 4. This method does not have the limitations of the ELL and ELL updating methods mentioned above, as will be elaborated in the Sections that follow. Moreover, this method also has advantages over the GLSMs, which are also an extension of the SPREE method, in terms of its applicability to Third World countries poverty data, and ease of implementation for large multiway tables.

In Section 5.2 the SPREE method is illustrated as an updating method followed by the description of the extended SPREE (ESPREE) method (Section 5.3). Details of the different steps in fitting the ESPREE model are described in Section 5.4 followed by a discussion of issues on fitting the ESPREE model when there are zero frequencies (Section 5.5). A discussion of the ESPREE method in comparison with the classical SPREE, GLSMs, and ELL-based updating methods is presented in Section 5.6. The Chapter summary is presented in Section 5.7.

5.2 SPREE as an Updating Method

The SPREE method as described in Chapter 4 can be implemented by fitting a generalized linear model (GLM) as follows,

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \tag{5.1}$$

to both the census and survey type data, where $g()$ is the log function, $\boldsymbol{\mu}$ is the expected value of the vector of the dependent variable, which for poverty estimation could be the number of households (which we denote here by $\boldsymbol{Y}$ for the survey period and $\tilde{\boldsymbol{Z}}$ for the census period) cross-classified by poverty status, province and other related variables. $\boldsymbol{X}$ is the design or model matrix corresponding to the explanatory variables.

In order to view SPREE as an updating technique, we emphasize the different time periods for the survey and census data by changing some notation in the SPREE model presented in the previous Chapter. The census model given in (4.7) from Section 4.3.1 is now as follows:

$$g(\boldsymbol{\mu}^{\tilde{Z}}) = \boldsymbol{X}_{1,t_0}\boldsymbol{\beta}_{1,t_0} + \boldsymbol{X}_{2,t_0}\boldsymbol{\beta}_{2,t_0} \tag{5.2}$$

with the subscript t_0 indicating that the data comes from an earlier period, while the survey model given earlier in (4.5) is now,

$$g(\boldsymbol{\mu}^{Y}) = \boldsymbol{X}_{1,t_1}\boldsymbol{\beta}_{1,t_1} + \boldsymbol{X}_{2,t_1}\boldsymbol{\beta}_{2,t_1} \tag{5.3}$$

with the subscript t_1 indicating that the data comes from a more recent period. We note that $\boldsymbol{X}_{1,t_0} = \boldsymbol{X}_{1,t_1} = \boldsymbol{X}_1$ and $\boldsymbol{X}_{2,t_0} = \boldsymbol{X}_{2,t_1} = \boldsymbol{X}_2$ are the partition of the design matrix corresponding to the partition of the parameter vector $\boldsymbol{\beta}$ in model (5.1), which is $\boldsymbol{\beta}_{1,t_0}$ and $\boldsymbol{\beta}_{2,t_0}$ for the census model, $\boldsymbol{\beta}_{1,t_1}$ and $\boldsymbol{\beta}_{2,t_1}$ for the survey model. As pointed out in Chapter 4, the first term in (5.3) generally represents the main effects and/or lower order parameters which can be accurately estimated from the survey data while the second partition represent the higher order terms which cannot be accurately estimated from the survey.

Considering models (5.2) and (5.3), SPREE is equivalent to fitting model (5.2) and then some of the lower order parameters in the model (first partition) are adjusted or updated in line with the most recent information available from the survey data while the higher order parameters (second partition), for which new information from the survey is not available, remains the same, i.e., $\boldsymbol{\beta}_{2,t_0} = \boldsymbol{\beta}_{2,t_1} = \boldsymbol{\beta}_2$ by assumption. In this way SPREE is used to generate updated small area estimates. This process is also equivalent to fitting model (5.3) and then equating the higher order parameters (second partition) to the values generated from the census model (5.2), that is:

$$g(\boldsymbol{\mu}^Y) = \boldsymbol{X}_{1,t_1}\boldsymbol{\beta}_{1,t_1} + \boldsymbol{X}_2\boldsymbol{\beta}_2 \tag{5.4}$$

5.3 The Extended SPREE Updating Model

As illustrated in the previous Section, the conventional SPREE method can be used for generating updated small area estimates. However, as mentioned in Section 4.6, one of the major limitations of the SPREE method is the assumption that $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0} = \boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$. To clarify this issue further, we consider the survey model:

$$g(\boldsymbol{\mu}^Y) = \boldsymbol{X}_1\boldsymbol{\beta}_{1,t_1} + \boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$$

Basically, in SPREE, the second term of the model, $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$, is approximated by $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0}$ since $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$ cannot be estimated from the survey data alone.

The GLSM proposed by Zhang and Chambers (2004) was developed primarily to weaken the requirement that $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0} = \boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$. The authors set

$$\begin{aligned}\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1} =& \boldsymbol{X}_2(\boldsymbol{\beta}_{2,t_0} + B)\\ =& \boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0} + \boldsymbol{X}_2 B\end{aligned}$$

Note that here B denotes bias and under SPREE the assumption is that $\boldsymbol{X}_2 B = 0$. The GLSM framework deals with this bias by introducing a proportionality coefficient which re-scales $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0}$, that is

$$\begin{aligned}\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1} =& \boldsymbol{X}_2(\boldsymbol{\beta}_{2,t_0} + B)\\ =& \zeta\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0}\end{aligned}$$

In this thesis, instead of specifying a scalar parameter to correct what is interpreted as bias, as in the GLSM, we propose a different approach as follows:

$$\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1} = \boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0} + \boldsymbol{\gamma}_{t_1} \tag{5.5}$$

We note that the error term, $\boldsymbol{\gamma}_{t_1}$ is the difference between $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$ and $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0}$. Model (5.4) can now be modified and written as,

$$g(\boldsymbol{\mu}^Y) = \boldsymbol{X}_1\boldsymbol{\beta}_{1,t_1} + \boldsymbol{X}_2\boldsymbol{\beta}_{2,t_0} + \boldsymbol{\gamma}_{t_1} \tag{5.6}$$

Using the census data from an earlier period and the most recent survey data, updated small area estimates of poverty measures may be generated by using model (5.6). However, the usual situation in Third World countries is that information on either poverty measures or per capita income (the variable of interest for poverty estimation) is not available in the census. Hence, generation of small area estimates of poverty measures and its updates through conventional SPREE presented in the previous Section is not feasible. A modified SPREE approach, suggested by Purcell and Kish (1980) can be considered; this approach, as described in Section 4.5, is equivalent to fitting an unsaturated log-linear model to the census type data or association structure by forcing some of the interaction terms to zero. The appropriateness of this method however, depends on whether these interactions are actually negligible, otherwise this would result in an increase in the bias of the estimates generated.

Given the limitation on the data available for updating small area estimates of poverty measures, we propose the use of the "pseudo-census" data from t_0 composed of replicates of the "estimated census" generated from the "modified ELL" method (the ELL framework is used but the survey regression fitting procedure is using the GSR method) at time t_0 described in Chapter 2. This procedure is generally carried out by fitting a regression model to the survey data and then applying the fitted model to the census data on the assumption that the census and survey have been conducted at the same time period.

Using the pseudo-census data from time t_0, the estimation problem is considered in the context of a superpopulation, i.e. we assume that the pseudo-census data has a superpopulation that produces $\{\tilde{\boldsymbol{Z}}\}$. Note that in our application, we have different sets of pseudo-census data derived from using the bootstrap for the variance components for time t_0. Under the assumption that the pseudo-census data forms a superpopulation, the coefficients $\boldsymbol{\beta}_{1,t_0}$ and $\boldsymbol{\beta}_{2,t_0}$ of the census model are now considered random (if we only have one set of pseudo-census data, the coefficients $\boldsymbol{\beta}_{1,t_0}$ and $\boldsymbol{\beta}_{2,t_0}$ are fixed) with respect to the superpopulation and with expectation $\boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_0}]$ and $\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}]$, respectively. The appropriate census model for each pseudo-census replicate can be written as:

$$g(\boldsymbol{\mu}^{\tilde{Z}}) = \boldsymbol{X}_1\boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_0}] + \boldsymbol{X}_2\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] + \boldsymbol{u}_{t_0} \tag{5.7}$$

where $\boldsymbol{u}_{t_0}$ is a random error vector with respect to the superpopulation assumed for the census cross-classification and that $\boldsymbol{u}_{t_0} = \boldsymbol{X}_1\boldsymbol{u}_{1,t_0} + \boldsymbol{X}_2\boldsymbol{u}_{2,t_0}$. Note that an estimator of the expected value here is given by the average of the pseudo-census values, which is essentially what SPREE (rather than ESPREE) does in equation (5.2) which uses a slightly different notation.

For the survey data, we will also assume that a superpopulation exists for the census from which the survey data is drawn. As in the census model, the parameters $\boldsymbol{\beta}_{1,t_1}$ and $\boldsymbol{\beta}_{2,t_1}$ of the survey model are now random (they are assumed fixed in SPREE) with respect to the superpopulation and with expectation $\boldsymbol{\xi}E[\boldsymbol{\beta}_{1,t_1}]$ and $\boldsymbol{\xi}E[\boldsymbol{\beta}_{2,t_1}]$, respectively. E is the expectation related to the sampling design of the survey data, while $\boldsymbol{\xi}$ is the expectation related to the superpopulation assumed for the census. However, we could assume that $\boldsymbol{\xi}E[\boldsymbol{\beta}_{1,t_1}] = \boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_1}]$ and $\boldsymbol{\xi}E[\boldsymbol{\beta}_{2,t_1}] = \boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_1}]$ provided that the sample is unbiased and the sampling design is uninformative. Hence, the survey model will now be:

$$g(\boldsymbol{\mu}^{Y}) = \boldsymbol{X}_1\boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_1}] + \boldsymbol{X}_2\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_1}] + \boldsymbol{u}_{t_1} \tag{5.8}$$

where $\boldsymbol{u}_{t_1}$ is a random error vector for the survey model and $\boldsymbol{u}_{t_1} = \boldsymbol{X}_1\boldsymbol{u}_{1,t_1} + \boldsymbol{X}_2\boldsymbol{u}_{2,t_1}$. Again we note that the second term of this model cannot be estimated from the survey data alone. We recall equation (5.5) and note that, by analogy, replacing the quantities here for SPREE by their expected values for ESPREE and retaining a random error term $\boldsymbol{\gamma}_{t_1}$,

$$\boldsymbol{X}_2(\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_1}] - \boldsymbol{\beta}_{2,t_1}) = \boldsymbol{X}_2(\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] - \boldsymbol{\beta}_{2,t_0}) + \boldsymbol{\gamma}_{t_1}$$

so that model (5.8) can now be written as:

$$\begin{aligned} g(\boldsymbol{\mu}^{Y}) =& \boldsymbol{X}_1\boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_1}] + \boldsymbol{X}_1\boldsymbol{u}_{1,t_1} + \boldsymbol{X}_2\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] + \boldsymbol{\gamma}_{t_1} + \boldsymbol{X}_2\boldsymbol{u}_{2,t_1} \\ =& \boldsymbol{X}_1\boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_1}] + \boldsymbol{X}_2\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] + \boldsymbol{\gamma}_{t_1} + \boldsymbol{u}_{t_1} \end{aligned}$$

where $\boldsymbol{u}_{t_1}$ is as defined in equation (5.8) above and $\boldsymbol{\gamma}_{t_1} = \boldsymbol{X}_2(\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_1}] - \boldsymbol{\beta}_{2,t_1}) - \boldsymbol{X}_2(\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] - \boldsymbol{\beta}_{2,t_0})$. Hence our proposed updating model which we call *Extended*

SPREE (ESPREE) is as follows,

$$g(\boldsymbol{\mu}^Y) = \boldsymbol{X}_1\boldsymbol{\xi}[\boldsymbol{\beta}_{1,t_1}] + \boldsymbol{X}_2\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] + \boldsymbol{\epsilon}_{t_1} \tag{5.9}$$

where $\boldsymbol{\epsilon}_{t_1} = \boldsymbol{u}_{t_1} + \boldsymbol{\gamma}_{t_1}$. Under the superpopulation model, the bias, $\boldsymbol{\gamma}_{t_1}$, is incorporated into a superpopulation variance term. This is based on the assumption that relative to the superpopulation, this bias has expected value zero. Nevertheless, it affects overall mean square error through the superpopulation variance.

We emphasize that one of the important assumptions of the ESPREE model is that that the initial regression-based pseudo-census cross-classification is an unbiased (or less biased) estimator of the actual cross-classification in the most recent period t_1. That is, under (5.5) we assume that $\boldsymbol{X}_2\boldsymbol{\xi}[\boldsymbol{\beta}_{2,t_0}] + \boldsymbol{\gamma}_{t_1}$ is an unbiased (or less biased) estimator of $\boldsymbol{X}_2\boldsymbol{\beta}_{2,t_1}$. In addition, we also assume that there is no net migration in between the census and survey periods. This assumption for the ESPREE method is similar to one of the inherent assumptions of the ELL-based updating method described in Chapter 3.

5.4 ESPREE Modelling Procedure

The fitting algorithm of the ESPREE updating method is adapted from the model fitting procedure for classical SPREE in Section 4.3.3 which could be summarized in the following steps:

1) Use the ELL method or regression-based approach to generate replicates of the pseudo-census cross-classification ($\tilde{Z}$) for the period t_0. This step requires survey data at t_0 containing information on income/expenditure and auxiliary variables and census data containing information on auxiliary variables. This step basically follows the ELL procedure described in Chapter 2.

2) Choose a suitable replication method for generating the distribution of the survey margins at time t_1.

3) Use the replicates of survey margins from step 2 to generate the pseudo-counts ($\hat{\hat{Y}}$) by a closed form formula or IPF depending on the structure of the design matrix $\boldsymbol{X_1}$.

These pseudo-count replicates should satisfy or sum up to the selected replicates of the survey margins.

4) Scale the mean of the replicates of the pseudo-census from step 1 to agree with the survey estimates of the margin counts in the period t_1. This can be accomplished by fitting the loglinear model in the form of equation (4.9) from Section 4.3.3 which requires the pseudo-count replicates generated from step 3.

For example if we are dealing with a two-way cross-classification and wanted to generate estimate of Y_{ab}, we will have:

$$\log\mu^{Y}_{ab} = \mathbf{X}_1\Delta\boldsymbol{\beta}_1 + \log\mu^{\tilde{Z}}_{ab} + u_{t_1}$$

This model is fitted with $\log\mu^{\tilde{Z}}_{ab}$ =offset (the log of the census data). We note that $g(\boldsymbol{\mu}^Y) = \log\mu^{Y}_{ab}$, u_{t_1} is a random error from the pseudo-count replicates and the other parameters are as defined earlier.

5) Fit the ESPREE model again in the form of equation (4.9) as in Step 4 above. However, the pseudo-census cross-classification this time will be variable (the census cross-classification is not fixed under ESPREE), that is we will use the actual replicates of the pseudo-census data instead of the mean of the pseudo-census replicates. The pseudo-counts on the other hand would be fixed, unlike in Step 4 that it was the one that was varying. The pseudo-counts value is fixed at the overall survey-based estimate of the margins.

6) The updated small area estimate is generated by using the model fitted in Step 4. The cross-classifications involved are the mean of the replicates of the pseudo-census and the pseudo-counts computed from the overall survey margins. The variance of the estimate is computed by adding the estimated variances generated from Steps 4 and 5. Details of the different variance estimation techniques are presented in Chapter 6.

5.5 Fitting ESPREE Models with Zero Frequencies

As with any statistical method that involves loglinear models, one of the issues encountered in using the ESPREE updating method for small area estimates is zero frequencies in the contingency table used for updating. Zero cells for census type data and survey type data complicates the estimation process because the log of the odds ratios that contribute to the estimates of the parameter $\boldsymbol{\beta}$ then contains zero in the numerator or denominator or both.

The occurrence of empty cells can be classified into two types based on the mechanism that causes this to happen: the first type arises due to the small probability of the event occurring that corresponds to that cell and is called a random zero; the second type of empty cell is one that has a priori a value of zero and hence is considered a non-random occurrence. This type of empty cell is called a structural zero. This corresponds to cells for which it is impossible for the combination of levels of factors to occur (Simonoff, 2003).

Under the ESPREE method, the census data, especially the pseudo-census values are assumed to be realizations of some underlying superpopulation process. Hence, the empty cells under ESPREE are considered as random zeros. Grizzle et al. (1969) advocate replacing these zeros with a small positive value and as pointed out by Purcell and Kish (1979) it has become a generally acceptable practice to add 1/2 to all zero cells (random zeros), but they added that this idea however has not been fully investigated. In fitting ESPREE model, we did a simulation to assess the effect of using different values for the empty cells, for example 0.001, 0.0001, and others. It was observed that the choice of value had no evident effect on the resulting small area estimates. In the application of the ESPREE method to the Philippine data, 0.0001 is used.

For the survey data, having empty cells is not a problem when using only a few survey margins. Provided that all the survey margins are positive, then all the pseudo-counts that will be used for estimation of parameters are also greater than zero. In cases where there are also zero pseudo-counts, a similar approach used to deal with zero cells in the census could be adopted. This means that techniques such as those in Haslett (1990) are unnecessary.

5.6 Comparison of ESPREE with Selected Methods

- **ESPREE vs Classical SPREE**

One of the main differences of the ESPREE method from the classical SPREE is that under ESPREE the association structure is assumed stochastic while it is assumed fixed under SPREE. Hence ESPREE has two sources of variation as specified in model (5.9) - the survey margins as well as the census data by using the pseudo-census data which are assumed to be random samples from the superpopulation. In the application of the ESPREE method in Chapter 7 using the Philippine data, the pseudo-census data are bootstrap estimates generated from the modified ELL method.

The classical SPREE method generates the small area estimates by using the IPF method which implicitly fits a generalized linear model, specifically a loglinear model. Estimation is done by adjusting the census data (in a contingency table) to satisfy the survey margins. This method can be very tedious and complicated when dealing with so many auxiliary variables which is generally the case in real world application for example in poverty estimation. The ESPREE method on the other hand, explicitly fits the generalized linear model to the survey margins and uses the census data as an offset. Directly fitting the loglinear models allows estimation of the model parameters and their estimated covariance matrix.

- **ESPREE vs GLSM**

The primary aim of using the GLSM is reduction of the bias coming from the assumption that the association structure does not change from the census period to the survey period. The proportionality coefficient is introduced to account for the changes in the association structure, hence to reduce the bias. However as pointed out and illustrated in Section 4.6, this coefficient can be highly influenced by some groups of loglinear parameters. Under the ESPREE procedure, bias is reduced by improving the loglinear models formulated by incorporating important auxiliary variables so that only valid relationships with the variable of interest are considered or by ensuring that spurious relationships in the model are minimized. Model improvement could sometimes mean using a larger number of auxiliary variables. Increasing the number of variables under the GLSM procedure could lead to an increase in the number of less stable loglinear parameter estimates that could heavily influence the proportionality coefficient.

The GLSM also implicitly assumes that the population or small area counts (during the non-census or intercensal period) are known in the example shown by Zhang and Chambers (2004) to illustrate the GLSM theory. This is not generally the case for the poverty updating problem in Third World countries, during non-census or intercensal years the population for the small areas of interest are not known or reliable estimates are not available. Under the ESPREE method, only the survey margins for the auxiliary variables and the variable of interest are assumed known.

- **ESPREE vs ELL-based Updating Method**

The ELL-based updating method requires either a panel survey data set or the existence of time-invariant auxiliary variables in cross-sectional survey data. Unlike cross-sectional survey data, panel survey data (for poverty estimation) are not generally available in Third World countries. While cross-sectional survey data are available, the proponents of the ELL-based updating method have not developed a proper statistical method to assess time-invariance of auxiliary variables.

The ESPREE method on the other hand, does not have the limiting assumptions on using time-invariant variables. Based on the description of the ESPREE method presented in the previous Sections, the ESPREE method uses auxiliary variables in such a way that structural changes from the census period to the most recent period are accounted for. In addition, the ESPREE method allows for the use of more useful auxiliary variables that do change in relation to changes in poverty, e.g. housing quality. Although this variable is considered to be a time invariant variable in the application of the ELL method in the Philippines. The use of time-invariant variables can increase the bias in the generated small area estimates, as will be shown in the comparison of the results of the ESPREE and ELL methods in Chapter 7.

Overall, among the four methods that can be used for updating small area estimates of poverty in Third World countries, the ESPREE method appears to be the best available method so far in terms of theoretical assumptions, data requirements, and implementation.

5.7 Summary

Details of the proposed small area updating method, which is an extension of the classical SPREE method, are presented in this Chapter. The step-by-step implementation of the method was also described. Various features of the ESPREE method were discussed and compared with other methods that can be used for updating small area estimates of poverty statistics in Third World countries. As mentioned, among the different methods available, it appears that the ESPREE method is the best method so far that can be used to generate updated small area estimates of poverty statistics in third world countries.

The different steps of implementation of the ESPREE method have been discussed in this Chapter, except for the generation of estimated variance of small area estimates. The next Chapter is devoted to the discussion of different variance estimation procedures that can be employed to generate the estimated variance of updated small area estimates.

Chapter 6

Variance Estimation

6.1 Introduction

Estimated variances have an important role in the small area estimation of poverty measures in a government's decision on aid allocation, particularly in choosing priority areas. Estimated variances or standard errors provide information on the precision of the small area estimates and whether the observed differences in the estimates signify real differences. A flawed estimation procedure could result in underestimation or overestimation of the variances. Underestimated variances for example could lead to declaring that one local area has a different poverty incidence rate than another area. Hence, allocating more resources to one area when both areas have similar amount of aid needed, thereby furthering economic inequality in local areas. The estimation procedure for deriving standard errors for the small area estimates of poverty measures should therefore be chosen carefully.

In this Chapter, variance estimation procedures for the SPREE method and the proposed Extended SPREE (ESPREE) method are investigated. There are two approaches that are usually employed for variance estimation when dealing with nonlinear estimators (e.g., poverty incidence and other type of ratios, differences of ratios, regression and correlation coefficients)- *linearization* also known as the *Taylor Expansion* or *delta* method, and *replication* methods. The *linearization* method involves approximating the nonlinear estimator by a linear function of the observation through Taylor's theorem, then applying the variance formula (appropriate to the sampling design) to the linear approximation. The *replication* method on the other hand, involves the calculation of the estimate of interest from the full sample as well as from a number of subsamples. The variation among the subsample estimates is used to derive the estimate of the variance of the full sample. There are various ways of generating the subsamples under the replication method, see for example Wolter (1985), Lohr (1999), Judkins (1990), Krewski and Rao (1981) and many others. The common methods being the balanced repeated replicates (BRR), jackknife and bootstrap method. Implementation of these methods for SPREE and ESPREE are discussed in the Sections that follow.

The discussion of the variance estimation procedure here is not an attempt to discuss and compare all the possible variance estimation procedures. The estimation techniques presented are selected based on the applicability to ESPREE and the available data for poverty estimation in Third World countries.

6.2 Variance Estimation for SPREE

Purcell (1979) discussed variance estimation for SPREE using linearization and replication (BRR and jackknife) methods. The estimation procedures for SPREE are all based on the assumption that the *census* data (association structure) is fixed and that the only source of variation of the estimates is the set of the most recent *survey margins* (allocation structure). We note that in this Chapter, as in Chapters 4 and 5, census data is used interchangeably with association structure and survey margins with allocation structure. The following notations are used for convenience and simplicity of exposition in discussing the different variance estimation procedures: $\hat{\mathbf{p}}$ is the column vector of the required cell estimates, i.e., $\hat{\mathbf{p}} = (\hat{p}_{111}, ..., \hat{p}_{ABC})'$ with dimensions $ABC \times 1$ where ABC is the total number of cells; $\mathbf{p}^*$ is the vector containing the allocation structure (elements are $\hat{p}_{.bc}$) and $\boldsymbol{\pi}$ is the vector with dimensions similar to $\hat{\mathbf{p}}$ containing the association structure (relative cell frequencies π_{abc} established in the most recent census year) which is assumed to be fixed and not subject to error under SPREE. However this is not the case for ESPREE as mentioned earlier and is elaborated further in the next Section.

Under the linearization method, Purcell (1979) expressed the relationship between the small area estimates and the allocation and association structures as follows:

$$\hat{\mathbf{p}} = F(\mathbf{p}^*; \boldsymbol{\pi}) \tag{6.1}$$

where F is the function defined by the IPF procedure. It is assumed that the only stochastic elements in $\hat{\mathbf{p}}$ are the most recent survey margins specified in the allocation structure $\mathbf{p}^*$ with its expected value denoted by $\boldsymbol{\vartheta}$, i.e., $E\{\mathbf{p}^*\} = \boldsymbol{\vartheta}$. Using the first order Taylor series approximation, the variance-covariance matrix of $\hat{\mathbf{p}}$ is approximated by:

$$V_{\hat{p}} = \left[\frac{\partial \hat{\mathbf{p}}}{\partial \boldsymbol{\vartheta}}\right] V_{\mathbf{p}^*} \left[\frac{\partial \hat{\mathbf{p}}}{\partial \boldsymbol{\vartheta}}\right]' \tag{6.2}$$

where $V_{\mathbf{p}^*} = E(\mathbf{p}^* - \boldsymbol{\vartheta})(\mathbf{p}^* - \boldsymbol{\vartheta})'$ and $\frac{\partial \hat{\mathbf{p}}}{\partial \boldsymbol{\vartheta}} = \frac{\partial (F(\mathbf{p}^*;\boldsymbol{\pi}))}{\partial \mathbf{p}^*} |_{\mathbf{p}^*=\boldsymbol{\vartheta}}$. See Purcell (1979) for details of the generation of specific terms of the matrix of partial derivatives.

For the replication methods, Purcell (1979) described the BRR and jackknife methods for SPREE. However there was no empirical investigation done as to the performance of these methods in generating the estimated variances for SPREE small area estimates. In the author's description of the BRR method for SPREE, the allocation structure was assumed to be estimated from a sampling design with two primary sampling units (PSUs) in each stratum. Generally, the BRR method can only be applied when we have two units in a stratum. In some applications however merging, dividing of units or pairing of strata are done in order to satisfy the requirements of the BRR procedure (see for examples (Lohr (1999); Wolter (1985)).

The jackknife method described for SPREE on the other hand considers not just one sampling design for the allocation structure. For allocation structures with only two PSUs in each stratum, Purcell (1979) described a method related to the BRR method. For sampling designs with more than two PSUs in a stratum, the *generalized jackknife estimator* proposed by Mellor (1973) was considered applicable for SPREE estimates. The different variance estimators for SPREE using BRR and jackknife methods are presented in Appendix C.

6.3 Variance Estimation Methods for Intercensal Small Area Estimates

A similar set of the variance estimation procedures presented above are proposed and described for ESPREE in this Section, but are appropriately modified to allow for the variation in the association structure (assumed fixed in SPREE). Under ESPREE, both the census and the survey margins are assumed variable. The bootstrap method is also presented since this method was used in the generation of the pseudo-census data from the small area estimation project conducted in the Philippines for the census year (Haslett and Jones, 2005), as pointed out in the previous Chapter, these are replicates of census data or the allocation structure. The data from the Philippines is used for the preliminary application of the ESPREE method. However, this should not be construed to suggest that the bootstrap method can only be used for the generation of the component of ESPREE estimated variance from the census data, it can also be used to generate the variance component from the survey margins.

6.3.1 Linearization Method

To derive the variance of the ESPREE estimates using the linearization method, we note that under the ESPREE method, the vector of cell estimates $\hat{\mathbf{p}}$ has now stochastic elements from both the allocation and the association structure. We let $\mathbf{\Pi}$ be the expected value of $\boldsymbol{\pi}$ ($E(\boldsymbol{\pi}) = \mathbf{\Pi}$)and as defined earlier, $\boldsymbol{\vartheta}$ is the expected value of $\mathbf{p}^*$ ($E(\mathbf{p}^*) = \boldsymbol{\vartheta}$). Using the first order Taylor series, we will have

$$\begin{aligned}\hat{\mathbf{p}} &= F(\mathbf{p}^*;\boldsymbol{\pi}) \\ &\approx F(\boldsymbol{Z};\mathbf{\Pi}) + \frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \mathbf{p}^*}(\mathbf{p}^* - \boldsymbol{\vartheta}) + \frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi} - \mathbf{\Pi})\end{aligned} \tag{6.3}$$

where F is defined by the generalized linear model. Assuming independence of $\mathbf{p}^*$ and $\boldsymbol{\pi}$, the variance-covariance matrix of $\mathbf{p}$ is then approximated by the variance-covariance matrix of the linear function (6.3), that is

$$\begin{aligned}V(\hat{\mathbf{p}}) =& E\left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \mathbf{p}^*}\right](\mathbf{p}^* - \boldsymbol{\vartheta})(\mathbf{p}^* - \boldsymbol{\vartheta})'\left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \mathbf{p}^*}\right]' \\ &+ E\left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right](\boldsymbol{\pi} - \mathbf{\Pi})(\boldsymbol{\pi} - \mathbf{\Pi})'\left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right]' \\ =& \left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \mathbf{p}^*}\right]V(\mathbf{p}^*)\left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \mathbf{p}^*}\right]' + \left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right]V(\boldsymbol{\pi})\left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right]'\end{aligned} \tag{6.4}$$

where $V(\mathbf{p}^*)$ and $V(\boldsymbol{\pi})$ are the covariance matrix for $\mathbf{p}^*$ and $\boldsymbol{\pi}$, respectively. That is, $V(\mathbf{p}^*) = E[(\mathbf{p}^* - \boldsymbol{\vartheta})(\mathbf{p}^* - \boldsymbol{\vartheta})']$ and $V(\boldsymbol{\pi}) = E[(\boldsymbol{\pi} - \mathbf{\Pi})(\boldsymbol{\pi} - \mathbf{\Pi})']$. We note that $V(\mathbf{p}^*)$ can be obtained either directly or by using any of the replication methods. While it is generally assumed that census is fixed, in some cases as will be shown in Section 6.3.2, $V(\boldsymbol{\pi})$ can be estimated using replication methods. It can be observed from equation (6.4) that under ESPREE the estimated variance is the sum of the variability from the association structure and the allocation structure.

A related result for the variance derived from linearization method could be attained by the approach proposed by Haslett et al. (1998). The approach formulated was based on the mean square error formula

$$[M\hat{S}E(\hat{\mathbf{p}}|\boldsymbol{\pi})] = V(\hat{\mathbf{p}}|\boldsymbol{\pi}) + (\hat{E}(\hat{\mathbf{p}}|\boldsymbol{\pi}) - \hat{\mathbf{P}})(\hat{E}(\hat{\mathbf{p}}|\boldsymbol{\pi}) - \hat{\mathbf{P}})' \tag{6.5}$$

where $\hat{\mathbf{P}}$ denote the set of cell estimates based on their long term averages. The estimated mean square error gives an estimate of the joint design/model (superpopulation) variance by treating both terms in the equation as estimates of conditional variances, see (Noble, 2003) for detailed derivation and proof. The first term on the right hand side of the equation can be considered to be the variability from the allocation structure (survey) and the second term from the association structure (census) as in the variance formula in (6.4).

One of the limitations of the linearization method is that the derivation of the partial derivatives could be complicated for some estimators. Purcell (1979) has derived the partial derivatives in the first component of the estimated variance in equation (6.4) as follows:

$$\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \mathbf{p}^*}|_{\mathbf{p}^*=\boldsymbol{\vartheta}} = \mathfrak{D}_p\tilde{\mathbf{A}}'(\tilde{\mathbf{A}}\mathfrak{D}_p\tilde{\mathbf{A}}')^- \tag{6.6}$$

where $\tilde{\mathbf{A}}$ is a matrix defined by Purcell (1979) containing the coefficients whose rows generate the required marginal relative frequencies that define the known allocation structure, i.e. $\tilde{\mathbf{A}}\hat{\mathbf{p}} = \mathbf{p}^*$ and $\mathfrak{D}_p$ is a diagonal matrix with the vector $\hat{\mathbf{p}}$ on the diagonal i.e. $\mathfrak{D}_p = \text{diag}\{\hat{\mathbf{p}}\}$.

As stated previously, under the ESPREE method the census or association structure is assumed stochastic. Hence we need to generate the partial derivatives with respect to the association structure, i.e. the partial derivatives in the second component of the right hand side of the variance formula: $\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}|_{\boldsymbol{\pi}=\boldsymbol{\Pi}}$. Given the set of partial derivatives above for the first component from Purcell (1979), for convenience, we use a similar approach to derive the second set of partial derivatives. We recall from Section 4.2.1 that if we have the available allocation structure consisting only of the BC margins then as given in equation (4.1), we have:

$$\hat{p}_{abc} = (\pi_{abc}/\pi_{.bc})\hat{p}_{.bc} \tag{6.7}$$

By taking logarithms, equation (6.7) could be expressed as:

$$\text{log}\hat{p}_{abc} = \text{log}\pi_{abc} + \lambda_{bc} \tag{6.8}$$

where $\lambda_{bc} = \text{log}\hat{p}_{.bc} - \text{log}\pi_{.bc}$. Equation (6.8) can be expressed in matrix notation as

$$\text{log}\hat{\mathbf{p}} = \text{log}\boldsymbol{\pi} + \tilde{\mathbf{A}}'\boldsymbol{\lambda} \tag{6.9}$$

where $\hat{\mathbf{p}}$ is an $ABC \times 1$ column vector of $\hat{p}_{abc}$, $\boldsymbol{\pi}$ is an $ABC \times 1$ column vector of π_{abc}, $\boldsymbol{\lambda}$ is a $BC \times 1$ column vector, and $\tilde{\mathbf{A}}$ is a $BC \times ABC$ matrix comprising of coefficients whose rows generate the required marginal relative frequencies that define the known allocation structure, that is,

$$\tilde{\mathbf{A}}\hat{\mathbf{p}} = \mathbf{p}^* \tag{6.10}$$

where $\mathbf{p}^*$ is a column vector of $\hat{p}_{.bc}$. Multiplying both sides of equation (6.9) with a matrix say $\mathbf{K}$ which is an orthocomplement of $\tilde{\mathbf{A}}$ (i.e. $\mathbf{K}\tilde{\mathbf{A}}' = \mathbf{0}$), we will have the interaction terms that are preserved or the condition equivalent to the assumption for the second term on the right hand side of equation (4.8) from Section 4.3.1, and here it will be

$$\mathbf{K}\text{log}\hat{\mathbf{p}} = \mathbf{K}\text{log}\boldsymbol{\pi} \tag{6.11}$$

In these notations, the SPREE and the ESPREE method are then characterized by equations (6.10) and (6.11). Again we note that under SPREE, $\boldsymbol{\pi}$ is assumed fixed but under ESPREE $\boldsymbol{\pi}$ is assumed stochastic. Following Purcell (1979) we evaluate the partial derivatives $\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}|_{\boldsymbol{\pi}=\boldsymbol{\Pi}}$ by differentiating both sides of equations (6.10) and (6.11) with respect to $\boldsymbol{\pi}$ and then evaluate the results at $\boldsymbol{\pi} = \boldsymbol{\Pi}$.

Here, we assume that the matrix $\tilde{\mathbf{A}}$ is of full row rank to derive the required partial derivatives. Taking the derivatives with respect to $\boldsymbol{\pi}$ we will have:

$$\begin{bmatrix} \tilde{\mathbf{A}} \\ \hdashline \mathbf{K}\mathfrak{D}_p^{-1} \end{bmatrix} \left[\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}|_{\boldsymbol{\pi}=\boldsymbol{\Pi}} \right] = \begin{bmatrix} \mathbf{0} \\ \hdashline \mathbf{K}\mathfrak{D}_\pi^{-1} \end{bmatrix}$$

where $\mathfrak{D}_\pi$ is a diagonal matrix with the vector $\boldsymbol{\pi}$ in the diagonal, i.e. $\mathfrak{D}_\pi = \text{diag}\{\boldsymbol{\pi}\}$ so that

$$\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}|_{\boldsymbol{\pi}=\boldsymbol{\Pi}} = \begin{bmatrix} \tilde{\mathbf{A}} \\ \hdashline \mathbf{K}\mathfrak{D}_p^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{0} \\ \hdashline \mathbf{K}\mathfrak{D}_\pi^{-1} \end{bmatrix}$$

Purcell (1979) illustrated that

$$\left[\begin{array}{c}\tilde{\mathbf{A}}\\ \hdashline \mathbf{K}\mathfrak{D}_p^{-1}\end{array}\right]^{-} = \left[\begin{array}{c|c}\mathfrak{D}_p\tilde{\mathbf{A}}'(\tilde{\mathbf{A}}\mathfrak{D}_p\tilde{\mathbf{A}}')^{-1} & \mathbf{K}'(\mathbf{K}\mathfrak{D}_p\mathbf{K}')^{-1}\end{array}\right]$$

is a right (generalized) inverse of the first term (assuming the term has full row rank since it is a non-square matrix) on the right side of the partial derivative of F, hence

$$\frac{\partial F(\mathbf{p}^*;\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \,|_{\boldsymbol{\pi}=\boldsymbol{\Pi}} = \mathbf{K}'(\mathbf{K}\mathfrak{D}_p\mathbf{K}')^{-1}\mathbf{K}\mathfrak{D}_{\pi}^{-1} \tag{6.12}$$

Therefore, the variance formula for the ESPREE estimates is as follows:

$$\begin{aligned} V(\hat{\mathbf{p}}) = & \left[\mathfrak{D}_p\tilde{\mathbf{A}}'(\tilde{\mathbf{A}}\mathfrak{D}_p\tilde{\mathbf{A}}')^{-1}\right] V(\mathbf{p}^*) \left[\mathfrak{D}_p\tilde{\mathbf{A}}'(\tilde{\mathbf{A}}\mathfrak{D}_p\tilde{\mathbf{A}}')^{-1}\right]' \\ & + \left[\mathbf{K}'(\mathbf{K}\mathfrak{D}_p\mathbf{K}')^{-1}\mathbf{K}\mathfrak{D}_{\pi}^{-1}\right] V(\boldsymbol{\pi}) \left[\mathbf{K}'(\mathbf{K}\mathfrak{D}_p\mathbf{K}')^{-1}\mathbf{K}\mathfrak{D}_{\pi}^{-1}\right]' \end{aligned} \tag{6.13}$$

6.3.2 Replication Methods

There are many replication methods available, however we only consider here three methods that are more likely to be useful for intercensal updating of small area estimates (especially for poverty measures), these are the BRR, jackknife and bootstrap methods. We will describe these methods as applied to intercensal updating below. As pointed out in Section 6.3.1, the variance of intercensal estimates is the sum of the variability from the census and the survey margins. The variability from the census, the second term in equation (6.4), can be obtained through any of the replication methods by having the survey margins fixed and varying the values of the census data. On the other hand, the variability from the survey margins, the first term in equation (6.4) can be generated by having the census data fixed and varying the values of the survey margins. Different replication method can be used for the two different components.

The BRR Method

The standard BRR design assumes that a population of PSUs can be grouped into G strata with two PSUs selected from each stratum using with-replacement sampling. Then, $\tilde{H}$ replicate half-sample estimates can be formed by selecting one of the two

PSUs from each stratum, based on a Hadamard matrix (used in order to generate balanced sample), and then using only the selected PSU to estimate the parameter of interest. When sampling weights are involved, the weights of the selected PSU is doubled. In order to obtain a balanced set of replicates the number of replicates used needs to be a multiple of four greater than or equal to the number of strata. BRR techniques has been used for a long time in survey estimation, for more detailed description see for example Lohr (1999), Wolter (1985), and Kovar (1985).

To derive the first term of the variance of ESPREE estimates using BRR, we let $\hat{\mathbf{p}}$ be the full sample estimates based on the survey margins derived from the full sample and $\hat{\mathbf{p}}_{(\tilde{h})}$ the ESPREE estimates based on the survey margins estimated from the $\tilde{h}$th half-sample. The BRR estimator for the first term of the variance of ESPREE estimates $\hat{\mathbf{p}}$ is

$$V_1(\hat{\mathbf{p}})_{BRR} = (1/\tilde{H})\sum_{\tilde{h}=1}^{\tilde{H}}(\hat{\mathbf{p}}_{(\tilde{h})} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(\tilde{h})} - \hat{\mathbf{p}})' \tag{6.14}$$

where $\tilde{H}$ is the number of replicates. We note that the full sample ESPREE estimates $\hat{\mathbf{p}}$ and the half-sample estimates $\hat{\mathbf{p}}_{(\tilde{h})}$ here are derived with the census fixed, or in practice when only pseudo-census data is available (e.g., in the Philippines) the average of the replicates are used as the fixed value of the census.

A similar approach could be done to generate the second term of the variance equation with the $\hat{\mathbf{p}}$ fixed and derived from the full sample, while the census values (association structure) $\boldsymbol{\pi}$ are varying through BRR. The BRR estimator of the second term of the variance of the ESPREE estimate is

$$V_2(\hat{\mathbf{p}})_{BRR} = (1/\tilde{H}_c)\sum_{\tilde{h}_c=1}^{\tilde{H}_c}(\boldsymbol{\pi}_{(\tilde{h}_c)} - \boldsymbol{\pi})(\boldsymbol{\pi}_{(\tilde{h}_c)} - \boldsymbol{\pi})' \tag{6.15}$$

where $\boldsymbol{\pi}_{(\tilde{h}_c)}$ denotes the half-sample estimates and $\tilde{H}_c$ is the number of replicates formed from the census data. The estimated variance of the ESPREE estimates could then be computed as $V(\hat{\mathbf{p}})_{BRR} = V_1(\hat{\mathbf{p}})_{BRR} + V_2(\hat{\mathbf{p}})_{BRR}$. However, this is not always the case, the variance estimation procedure for one of the two terms may not necessarily be generated through BRR. Other estimation procedures could be combined with BRR. The BRR method is generally not convenient to use for the census as it would require a very large Hadamard matrix and the census would not have the required structure (e.g. two PSUs per stratum).

As pointed out above, the number of replicates should be at least equal to the number of strata in order to have a balanced set of replicates. In cases where the number of strata is large, it would be very expensive or time consuming to use all the replicates for variance estimation. One of the methods recommended is *partial balancing* of the half-samples (Wolter, 1985). Partial balancing is done by dividing the G strata into L groups with G/L strata in each group. A fully balanced set of $\tilde{H}$ half-samples is then specified for the first group and is repeated for the remaining L-1 groups. See Wolter (1985) for detailed description of the method.

The Jackknife Method

Another approach to estimating the variance of the ESPREE estimates is the jackknife procedure. This method was introduced by Quenouille (1949) to estimate bias of estimates for simple random sampling and is now one of the popular variance estimation techniques following the suggestion of Tukey (1958) that the recomputed statistics from the estimation of bias could also provide a non-parametric estimate of the variance. To explain the basic idea of jackknife estimation, say for a set of n observations we computed our full sample estimate $\hat{\mathbf{p}}$. To estimate the variance of $\hat{\mathbf{p}}$, replicates are formulated by excluding each observation from the data one at a time and a new estimate $\hat{\mathbf{p}}_i$ is computed from each set of replicates for the remaining $n-1$ observations. For the ESRPEE method, the jackknife estimate of the first component of the variance of $\hat{\mathbf{p}}$, i.e. the first component of equation (6.4) assuming the survey design used is simple random sampling, is then given by

$$V_1(\hat{\mathbf{p}})_J = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\mathbf{p}}_i - \hat{\mathbf{p}})(\hat{\mathbf{p}}_i - \hat{\mathbf{p}})' \tag{6.16}$$

This expression is similar to the ordinary simple random sampling variance formula for $\hat{\mathbf{p}}$ except that the factor is $(n-1)/n$ instead of $1/n$ or $1/(n-1)$ reflecting the fact that for each jackknife estimate $(n-1)/n$ of the original sample is retained.

For more complex survey design used in practice such as stratified and clustered designs, the extension of the jackknife method from the simple random sampling described above is not straightforward and there are various versions proposed. One of the jackknife variance estimators constructed by Krewski and Rao (1981) which requires less computational effort is as follows:

$$V_1(\hat{\mathbf{p}})_{J_1} = \sum_{g=1}^{G} m_g^{-1}(n_g - 1) \sum_{i=1}^{m_g} (\hat{\mathbf{p}}_{i(g)} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{i(g)} - \hat{\mathbf{p}})' \tag{6.17}$$

where only a random sample of size m_g ($< n_g$) of the n_g PSU's in the sample is used in stratum g. However, for efficiency considerations, some surveys are designed similar to BRR, where there are only two units that are selected per stratum - stratified half-sample designs. This case is similar to the 2003 Philippine survey data (see details in Section 7.4). The estimator in equation (6.17) reduces then to the jackknife variance estimator proposed by Frankel (1971) as presented by Purcell (1979), that is,

$$V_1(\hat{\mathbf{p}})_{J_2} = \sum_{g=1}^{G} (\hat{\mathbf{p}}_{(g)} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(g)} - \hat{\mathbf{p}})' \tag{6.18}$$

This estimator is a specific form of (6.17) where $n_g = 2$ and $m_g = 1$. In practical application this is done by leaving out one half-sample in the gth stratum but including twice the other selection in that stratum. As in the BRR, if sampling weights are involved, the sampling weight of the unit selected in a particular stratum is multiplied by 2. Kovar (1985) and Judkins (1990) have done some study comparing this estimator with other variance estimation procedures like BRR and linearization method. They found that the performance of this estimator is as good as other estimators and generates estimates that are similar to the other techniques asymptotically. An example comparing the results of the jackknife and linearization method is presented in Table C.1, Appendix C.

Similar to BRR, the usual problem for variance estimation using the jackknife method described here is when the number of strata is very large, which is very common in present day surveys. It is obvious from equation (6.18) that the number of replicates is required to be equal to the number of strata, i.e $\tilde{H} = G$ which could be very costly for large number of strata. For this case we propose that when the number of strata G is large, we can draw randomly a sample of the strata and compute the jackknife variance as follows:

$$V_1(\hat{\mathbf{p}})_{J_3} = (G/\breve{g}) \sum_{g=1}^{\breve{g}} (\hat{\mathbf{p}}_{(g)} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(g)} - \hat{\mathbf{p}})' \tag{6.19}$$

where $\breve{g}$ is the number of sample strata. This is equivalent to the use of a sample to estimate a population variance. This modified jackknife procedure has been tried on the generation of the estimated variance of survey margins for the Philippine data. Results showed that estimates were generally close to the linearization estimates. However, further investigation needs to be conducted to assess its performance under the ESPREE method.

Similar to the BRR, the jackknife method can also be used for either of the two components of the ESPREE estimated variance. Assuming that the jackknife procedure will be used to estimate the census and survey margins component of the variance of ESPREE estimates, we will have the second component as follows

$$V_2(\hat{\mathbf{p}})_{J_3} = (G/\breve{g}) \sum_{g=1}^{\breve{g}} (\boldsymbol{\pi}_{(g)} - \boldsymbol{\pi})(\boldsymbol{\pi}_{(g)} - \boldsymbol{\pi})' \tag{6.20}$$

so that $V(\mathbf{p})_{J_3} = V_1(\mathbf{p})_{J_3} + V_2(\mathbf{p})_{J_3}$.

In cases where there are varying number of clusters per stratum (i.e. some strata have large number of clusters/PSUs, some have very few), the implementation of the jackknife method described above can be very complicated. A jackknife technique more suitable for such data set is the delete-a-group jackknife (DAGJK) proposed by Kott (2001). As pointed out by the author this technique has no theoretical advantages over the jackknife method described above, but DAGJK offers computational advantages, is claimed to have simpler implementation and is easier to explain to external users of survey data.

DAGJK requires that the number of PSUs per stratum be large in all strata. However, in situations where there are only a few PSUs, the so called *extended DAGJK* can be used (Kott, 2001). Extended DAGJK allows estimation for various number of PSUs in a stratum - whether larger or smaller than the chosen number of random groups or replicates. As described by the author, the DAGJK procedure divides the (first-phase) sample into $\tilde{H}$ random groups (within each stratum) and then one group at a time is deleted from the sample, this is done by setting the sampling weight equal to zero when the PSU is a member of a particular group. The remaining PSUs in the stratum are used to compute the "replicate" estimate. The replicate estimate uses the adjusted weights. Weight adjustment is similar to the one shown in the stratified cluster design, however we use a more general factor, $n_g/(n_g - n_{g\tilde{h}})$ ($n_{g\tilde{h}}$ is the number of PSUs removed from the $\tilde{h}$th group in a particular stratum g). The variance estimate is then computed by taking the sum of the squared differences between the $\tilde{H}$ replicate estimates and the original estimate multiplied by $(\tilde{H}-1)/\tilde{H}$.

The key to extended DAGJK variance estimation technique is the development of the replicate weights ($w_{h(\tilde{h})}$), which are the sampling weights (of the hth element within a PSU or the PSU itself) adjusted to account for the sampling weight of the PSU (or the hth element within a PSU) that was "removed". Under this method, there are three situations considered:

1) number of PSUs is less than the number of random groups ($n_g < \tilde{H}$),

2) number of PSUs equal to the number of random groups ($n_g = \tilde{H}$) and

3) number of PSUs is greater than the number of random groups ($n_g > \tilde{H}$).

The corresponding recommended computation of replicate weights for each case are presented below, where $\tilde{H}$ is the number of replicates. Following Kott (2001), we let w_{gbh} be the sampling weight of element h in PSU b of stratum g, n_g and $\tilde{H}$ are as defined above, G is the number of strata, and $S_{g\tilde{h}}$ is the set of PSUs in stratum g and group $\tilde{h}$. The following are the recommended replicate sampling weights for the three cases, when $n_g < \tilde{H}$:

$$w_{gbh(\tilde{h})} = \begin{cases} w_{gbh} & \text{when } S_{g\tilde{h}} \text{ is empty} \\ w_{gbh}(1 - [n_g - 1]Z) & \text{when } b \text{ is in } S_{g\tilde{h}}, \text{ and} \\ w_{gbh}(1 + Z) & \text{otherwise.} \end{cases} \tag{6.21}$$

where $Z^2 = \tilde{H}/[(\tilde{H} - 1)n_g(n_g - 1)]$ and $w_{gbh(\tilde{h})}$ is the replicate weight (adjusted weight).

When $n_g > \tilde{H}$, the replicate sampling weights are similar to the DAGJK ("not extended" version) described above. Putting the replicate weights in the context of equation (6.21), we will have:

$$w_{gbh(\tilde{h})} = \begin{cases} 0 & \text{when } b \text{ is in } S_{g\tilde{h}}, \\ w_{gbh}(\frac{n_g}{n_g - n_{g\tilde{h}}}) & \text{otherwise.} \end{cases} \tag{6.22}$$

The replicate weights when $n_g = \tilde{H}$ is just similar to the stratified cluster design described earlier.

This method was tried on the 2000 survey data for estimating the variance of some survey margins. The variance estimates were very close to the variance estimates derived from the linearization method. This method also needs further investigation under the ESPREE method, however the design of the 2003 survey data has changed from the year 2000 design hence it was more suitable to use the BRR with partial balancing for the ESPREE method (see Section 7.4).

The Bootstrap Method

The bootstrap is the most recent replication technique of the three replication methods that is presented here. The development of this method came with the advancement in computer technology, this method needs a fast computer to simplify the usually complex calculations. Basically, the replicates in the bootstrap method are generated by drawing *with replacement* a sample of size n from the original sample (of size n). The process is repeated a large number of times, say 100 times, generating 100 replicates of the estimate (Efron and Tibshirani, 1993, suggested to use 25-200 replicates for estimation of the standard error). The generation of the bootstrap sample could be done either by drawing from the empirical distribution of the data (non-parametric bootstrap) or from a theoretical probability model (parametric bootstrap).

Presented here is an adaptation of the algorithm presented by Efron and Tibshirani (1993) to generate a bootstrap variance estimate. The algorithm starts with the selection of $\dot{S}$ independent bootstrap samples or replicates each consisting of n data points drawn with replacement from the original sample, the corresponding estimates computed from each of the replicates are as follows, ($\tilde{\mathbf{p}}^1$,$\tilde{\mathbf{p}}^2$,...,$\tilde{\mathbf{p}}^{\dot{S}}$) where $\dot{s} = 1, ..., \dot{S}$. The estimated bootstrap variance of the first component of the ESPREE estimated variance is computed as

$$V_1(\hat{\mathbf{p}})_{Bot} = \sum_{\dot{s}=1}^{\dot{S}} (\tilde{\mathbf{p}}^{\dot{s}} - \tilde{\mathbf{p}}(\cdot))(\tilde{\mathbf{p}}^{\dot{s}} - \tilde{\mathbf{p}}(\cdot))'/(\dot{S}-1) \tag{6.23}$$

where $\tilde{\mathbf{p}}(\cdot) = (\tilde{p}_{111}(\cdot), ..., \tilde{p}_{ABC}(\cdot))$ and $\tilde{p}_{abc}(\cdot) = \sum_{\dot{s}=1}^{\dot{S}} p_{abc}^{\dot{s}}/\dot{S}$ such that $a = 1, ..., A$, $b = 1, ..., B$ and $c = 1, ..., C$. Again this method can be used for variance estimation for either of the two components of the variance of the ESPREE estimates. If both components of the estimated variance are computed using a bootstrap method then the second component will be

$$V_2(\hat{\mathbf{p}})_{Bot} = \sum_{\dot{s}=1}^{\dot{S}} (\tilde{\boldsymbol{\pi}}^{\dot{s}} - \tilde{\boldsymbol{\pi}}(\cdot))(\tilde{\boldsymbol{\pi}}^{\dot{s}} - \tilde{\boldsymbol{\pi}}(\cdot))'/(\dot{S}-1) \tag{6.24}$$

where $\tilde{\boldsymbol{\pi}}(\cdot) = (\tilde{\pi}_{111}(\cdot), ..., \tilde{\pi}_{ABC}(\cdot))$ and $\tilde{\pi}_{abc}(\cdot) = \sum_{\dot{s}=1}^{\dot{S}} \tilde{\pi}_{abc}^{\dot{s}}/\dot{S}$. Assuming that the bootstrap method is used in computing both components of the estimated variance, we will have $V(\mathbf{p})_{Bot} = V_1(\mathbf{p})_{Bot} + V_2(\mathbf{p})_{Bot}$. Details of the application of the bootstrap method in the ELL method is presented in Section 2.3.3, the estimates generated (in the implementation in the Philippines) are in turn used as the the pseudo-census data (used as the association structure) in the application of the ESPREE method for the Philippine data which is given in Chapter 7.

6.4 Summary

The different methods of variance estimation - linearization and replication methods are discussed in this Chapter, both for the SPREE and the ESPREE methods. The major difference between the variance formula for SPREE and ESPREE is emphasized, i.e. the association structure (census) is assumed stochastic under the ESPREE method while it is assumed fixed under SPREE. Hence, the variance formula of the ESPREE estimates have two components - variability from the allocation structure and association structure.

The application of the ESPREE method to the Philippine data is presented in the next Chapter. Given that the number of strata in the Philippine survey data is very large, the BRR method with partial balancing is used for estimating the first component of the variance formula (variability from the allocation structure) of the ESPREE estimates. On the other hand, the second component (variability from the association structure) uses the bootstrap method since bootstrapping was used in the generation of the pseudo-census data which is used as the stochastic association structure for the ESPREE method. The pseudo-census data is an output of the poverty mapping project in the Philippines (Haslett and Jones, 2005) implemented by using the "modified ELL" method described in Section 5.4. See also Section 2.3.3 for details of the application of the bootstrap method under the ELL method.

Chapter 7

Application to the Philippine Data

7.1 Introduction

In this Chapter an application of the ESPREE updating method is illustrated using the survey and census data from the Philippines. The detailed description of the sources of data is presented in Section 7.2. Model formulation is described in Section 7.3 including variable selection and description of the available survey margins. Section 7.4 provides information on the estimation method used to generate the estimated variance of the updated small area estimates. The intercensal small area (municipal level) estimates and the accumulated estimates at the provincial and regional levels are presented including the estimates from the ELL-based updating method and the survey-based estimates in Section 7.5. This is followed by a discussion in Section 7.6 which emphasizes the advantages of the ESPREE updating method over the other existing methods.

7.2 The Data

The ESPREE method is applied to the Philippine national survey and census data. There are two sets of survey data used - Family Income and Expenditure Survey (FIES) and Labor Force Survey (LFS) from two periods, year 2000 and 2003 which are both conducted by the National Statistics Office (NSO). Basically, the ESPREE updating method only requires the 2003 survey and the 2000 census data. However, as pointed out in Chapter 5, as a preliminary we are going to form sets of pseudo-census data from the 2000 census and survey data. To generate the sets of pseudo-census data, information on the auxiliary variables from the survey conducted at the same time as the census are consequently necessary, so that it is useful to give a description of the survey data from the year 2000.

7.2.1 Survey Data

The FIES, as described in Section 2.4.4, collects information on family income and living expenditures as well as data on related variables affecting income and expenditure patterns and levels. The LFS on the other hand, collects data on employment and related information on demographic and socio-economic characteristics of the population over 15 years old. The reference period for LFS is the previous week which refers to the seven days preceding the date the data was gathered.

The FIES is conducted once every three years as a rider to the LFS which is conducted quarterly hence the two surveys are using the same survey design. The households included in the two surveys have a unique identifier that would allow the merging of the two data sets, allowing for a richer set of data on households. Data for FIES are gathered in two separate operations, each covering a half-year period in order to allow for seasonal patterns in income and expenditure. For FIES 2000 the interviews were conducted in July 2000, for the period 1 January to 30 June, and January 2001 for the period 1 July to 31 December. For the FIES 2003, interviews were conducted in same months as the 2000 FIES.

There are some differences in the survey design for the year 2000 and 2003. As discussed in Section 2.4.4, the 2000 survey used a multi-stage stratified random sampling method, wherein the barangays are the Primary Sampling Units (PSUs) which are stratified into urban or rural within each province and selected using systematic sampling with probability proportional to size. Large barangays are further divided into enumeration areas and subjected to further sampling before the final stage in which households are systematically sampled from the 1995 Population Census List of Households. This gave a total sample size of 41000 households. On the other hand, the 2003 survey design still used a multi-stage stratified random sampling method, however the definition of PSUs changed, the sampling domains and stratification variables were also modified and the sample households were selected from the 2000 Census of Population and Housing (CPH) described in detail in Section 7.2.2. The sample size of the 2003 survey consisted of 42094 households.

In the 2003 survey the PSUs were required to be a cluster of at least 500 households. There are barangays however with less than 500 households, hence some of the contiguous barangays within a municipality were then grouped together to form the required PSUs. The set of PSUs in each region were thereafter classified into certainty PSUs and non-certainty PSUs. The PSUs which are large, i.e., with selection probability of one, are considered certainty PSUs which are included outright in the sample. Each certainty PSU is considered a stratum. The non-certainty PSUs on the other hand are stratified within each province, highly urbanized city (HUC) or independent component city (ICC) by three socio-economic variables namely, proportion of strongly built houses, proportion of households engaged in agriculture and municipal per capita income and are selected using systematic sampling with probability proportional to size. The PSUs are further divided into enumeration areas which are also subjected to further sampling before the ultimate sampling stage in which households are selected from the 2000 CPH.

Presented in Table 7.1 is a summary of the coverage at various levels of FIES in the year 2000 and 2003. Interview non-response for the year 2000 was only 3.4 percent while 4.3 percent for the year 2003. Deterministic imputation was done to address item non-response, i.e., entry for a particular missing item is deduced from other items in the questionnaire. Note that some of the households in the two surveys are omitted either because of missing data or the urbanity or municipal codes did not match with the year 2000 codes. FIES and LFS are designed to give reliable estimates at regional level, based on the table below, the sample size is quite adequate for that purpose. However, the sample size for the province and municipality levels are not sufficient for reliable estimates. For the year 2000, about 25 percent of all municipalities are not sampled while about 30 percent for 2003. Even for the sampled municipalities the sample sizes become too small for direct estimation to be useful.

The PSUs sampled in FIES 2000 are derived from the 1995 census; hence they are not entirely compatible with those of the 2000 census. The 2003 sampled PSUs on the other hand are derived from the 2000 census and were supposed to be compatible with the 2000 census administrative boundaries. However, a new region was created in 2002 and some provinces have moved from one region to another or some municipalities have moved from one province to another, in addition, new barangays were also. These issues cause some difficulties in the merging of the survey and census data, we decided to use a consistent boundary assignment for the two sets of survey and census data. Since urbanity codes were not included in the 2003 survey data, the 2000 census boundary assignment were used and the urbanity classification in the 2000 census was applied to the 2003 survey data. In doing so, some of the sampled PSUs in 2003 were not included in the survey data that is used for the formulation of the ESPREE models. Despite these differences, PSUs used in the 2000 and 2003 surveys are very similar. Moreover, the 2000 Philippine Standard Geographic Code (PSGC) is used for consistency of codes for the census and survey data and simplicity of the implementation of the ESPREE method, hence the number of regions in the 2003 survey is still recorded as 16 instead of 17.

7.2.2 Census Data

In this study we are using the year 2000 Census of Population and Housing (CPH) conducted by the NSO. The CPH in the Philippines is conducted every ten years with a Census of Population every 5 years. A common questionnaire (short form) is given to all households, with an extended questionnaire (long form) completed by a random sample of about 10 percent of the population. The sampling design employed for this 10 percent sample is a systematic cluster design, with the sampled fraction being 100, 20, or 10 percent depending on the size of the municipality.

Table 7.1: Structure of the 2000 and 2003 survey data

	Region	Province/City	Municipality	Barangay	Household
2000					
FIES contains	16	83	1254	3366	39537
Mean num of households	2471.1	476.3	31.5	11.7	
Min num of households	1490	93	4	2	
Mean num of barangays	210.4	40.5	2.7		
Min num of barangays	127	8	1		
2003					
FIES contains	16	83	1134	2808	41759
Mean num of households	2609.9	503.1	36.8	14.9	
Min num of households	1467	16	2	1	
Mean num of barangays	175.5	33.8	2.5		
Min num of barangays	102	1	1		

The census was carried out from 1 May to 24 May with approximately 44000 enumerators. The population on census night (1 May) was declared to be 76.5 million. In conjunction with the enumeration of the population, a mapping operation was undertaken to update regional boundaries. Table 7.2 shows the coverage of the 10 percent census long form sample. Note that access to the administrative indicators of the long form was limited to regional, provincial and municipal level. For details of the variables (municipal averages) that were included in the set of explanatory variables used in the regression models for modeling income or expenditure see Haslett and Jones (2005).

Table 7.2: Structure of 10 percent long form census

	Region	Province/City	Municipality	Households
Long Form contains	16	83	1623	1511718
Mean number of households	94482	18213	931	
Minimum number of households	28618	1624	24	

Source: Haslett and Jones (2005)

The census data used is basically a set of replicates of the census data or is called in this research as a pseudo-census data. The pseudo-census data is the 100 bootstrap estimates of poverty status classification of the population. This is an output of the poverty mapping project conducted by the World Bank (WB) in collaboration with the National Statistical Coordination Board (NSCB) in the Philippines employing the modified ELL method as described in Chapter 2 (see also Haslett and Jones, 2005).

7.3 Model Formulation

7.3.1 The Auxiliary Variables

The set of variables (Table 7.3) used to illustrate the ESPREE method is a subset of the variables used in the collaborative project of the WB and NSCB mentioned above, see Haslett and Jones (2005) for a complete list and definition of variables. These variables have available information in both the 2000 census and the 2000 and 2003 survey data sets and are strongly correlated with household per capita income and hence the poverty status of members of the household.

Due to limited computer memory capacity for running the program for generating the estimates, the maximum number of variables considered is only six (6). With six auxiliary variables, the number of cells in the contingency table is already about 850,000. An example of a set of variables considered are as follows: urbanity (urban or rural), educational attainment (college education and no college education), type of house wall materials - strong, light and salvaged and household head gender (male or female). Fitting a regression model to the 2003 log of income per person with the variables above as explanatory variables yielded a multiple correlation coefficient of about 0.67 (or an $R^2 \simeq 0.5$), which is typical of many ELL applications. Hence for the Philippine data, even for fine-level household data, using six explanatory variables may be sufficient for generating reliable updated small area estimates of poverty measures which is a function of per capita income.

Under the ESPREE method however, a loglinear model is directly fitted to poverty status (poor and non-poor) hence, the relationship between the set of variables with poverty status needs to be checked. Two-way tables (cross-tabulation of poverty status with each of the auxiliary variables) including the chi-square values are presented

Table 7.3: List of auxiliary variables considered for ESPREE modelling

Variables	Definition
urb	1 if urban
famsize	number of persons in household
type_sing	1 if type of housing is single house
type_dup	1 if type of housing is duplex
type_mult	1 if type of housing is apartment/condominium/townhouse
type_cia	1 if type of housing is commercial/industrial/agriucltural building
type_oth	1 if type of housing is other
roof_strong	1 if roof is made of strong materials (galvanized iron/aluminum/tile/concrete/clay)
roof_light	1 if roof is made of light materials(cogon/nipa/anahaw)
roof_salvaged	1 if roof is made of salvaged materials(makeshift/improvised)
roof_oth	1 if roof is made of other materials
wall_strong	1 if wall is made of strong materials (galvanized iron/aluminum/tile/concrete/clay)
wall_light	1 if wall is made of light materials(cogon/nipa/anahaw)
wall_salvaged	1 if wall is made of salvaged materials(makeshift/improvised)
wall_oth	1 if wall is made of other materials
head_male	1 if head is male
no_spouse	1 if no spouse in family
all_noed	1 if theres a member of the family 10 years and over with no education
all_eled	1 if theres a member of the family 10 years and over with elementary education
all_hsed	1 if theres a member of the family 10 years and over with high school education
all_coed	1 if theres a member of the family 10 years and over with college education

in Appendix D. It can be observed that all the chi-square statistics are significant since all the chi-square values are greater than $\chi^2_{df=1,\alpha=.001}$ which is equal to 10.83.

Contrary to the key assumption of the ELL-based updating method presented in Chapter 3 that explanatory variables should be time invariant, the set of explanatory variables used for ESPREE are not required to be time invariant, which allows changes in the explanatory variables to explain change in the variable of interest or dependent variable which is poverty status. ESPREE allows for structural change in the model which is not possible under the ELL-based updating method. The ELL-based updating method fits a regression model using the 2003 survey data and applies this model to the census with the assumption that the regression model using the 2003 data still holds for the situation in the year 2000. Details of the variables used in the ELL-based updating method are presented in Chapter 3. In Table 7.4, the regression coefficients of the same model for log per capita income (lnincpp) for 2000 and 2003 are presented including the Z score, which is used to assess structural change between the two periods:

$$Z = \frac{\beta_{2003} - \beta_{2000}}{\sqrt{\mathrm{SE}^2_{2003} + \mathrm{SE}^2_{2000}}}$$

The Z score represents a measure of the standardized distance between the two sets of regression coefficients from different time periods, since there is little overlap at PSU level for the 2000 and 2003 surveys. It can be observed that the variables urbanity, wall type, and educational attainment have higher values of computed Z-score indicating larger discrepancies between regression coefficient estimates from different time periods. The significant difference between regression coefficients indicates structural change. The two way tables (in Appendix D) mentioned above could also strengthen the argument on structural change, the data from the two periods are tabulated side by side so that changes in the proportion can be illustrated clearly. It can be observed that majority of the tables show a significant change in proportion for poor and non-poor by auxiliary variables, for example, about 48 percent of the population in the rural area are non poor in the year 2000 while it was around 43 percent in the year 2003.

Table 7.4: Regression coefficient for 2000 and 2003 model for log per capita income

lnincpp	2000		2003		Z
	Coef.	SE	Coef.	SE	
urb	0.4146	0.0118	0.3837	0.0117	-1.8518
famsize	-0.1031	0.0022	-0.1006	0.002	0.8454
type_sing	-0.1688	0.1962	-0.1204	0.1054	0.2174
type_dup	-0.058	0.1952	-0.0382	0.107	0.0887
type_mult	0.0688	0.1982	0.0238	0.1068	-0.1995
type_cia	-0.0231	0.209	0.2706	0.1615	1.1121
roof_strong	0.0513	0.0168	0.0323	0.0158	-0.8246
roof_light	-0.2276	0.0181	-0.2021	0.017	1.0283
roof_salva d	-0.0447	0.055	-0.0715	0.0505	-0.359
wall_strong	0.2226	0.0159	0.2388	0.0142	0.7592
wall_light	-0.0898	0.0165	-0.1161	0.0156	-1.1582
wall_salva d	-0.1648	0.0544	-0.0109	0.0401	2.2768
head_male	-0.1077	0.0153	-0.125	0.0135	-0.8452
no_spouse	-0.0157	0.015	-0.0093	0.0126	0.3274
dom_help	0.8714	0.0394	0.8337	0.0276	-0.7827
all_noed	-0.1396	0.0145	-0.1586	0.0144	-0.9327
all_coed	0.5513	0.0094	0.5187	0.0082	-2.6048
all_eled	-0.0408	0.0082	-0.2089	0.0084	-14.3081
all_hsed	0.053	0.0097	0.0568	0.0081	0.2965
_cons	10.1079	0.1976	10.3377	0.1071	1.0225

7.3.2 The Survey Margins

Using the list of the variables in the previous Section, the corresponding survey margins along with standard deviations and coefficient of variation are presented in Table 7.5. The margins are national estimates from the combined 2003 FIES and LFS data. In some related applications of the formulation of the models, the margins are assumed to be independent (see Noble et al., 2002). However, for the case of the Philippine survey data, some of the margins are correlated as shown in Tables E.1 and E.2 in Appendix E. The correlations of the margins can be implicitly incorporated into the analysis by using any of the replication methods of variance estimation - jackknife, balanced repeated replicates (BRR) and others.

7.3.3 The Models

Various models were fitted and some of the models considered are presented in Table 7.6. Although alternatives exist, to assess the goodness of fit of the different models,

Table 7.5: Survey margins of the auxiliary variables

Variables	Margins	SD	CV
Non-poor	55,206,524	402,987	0.0073
Poor	23,502,040	309,823	0.0132
Wall_strong	45,575,436	398,044	0.0087
Wall_light	18,086,702	280,485	0.0155
Wall_salvaged	976,596	69,580	0.0712
Wall_others	14,069,827	321,318	0.0228
Rural	39,952,672	566,712	0.0142
Urban	38,755,892	583,793	0.0151
Female Headed HH	10,724,622	153,459	0.0143
Male Headed HH	67,983,944	396,484	0.0058
No HS	19,658,640	220,241	0.0112
HS	59,049,924	384,508	0.0065
Roof_strong	52,285,904	433,747	0.0083
Roof_light	15,349,936	247,819	0.0161
Roof_salvaged	712,601	56,741	0.0796
Roof_others	10,360,120	292,246	0.0282
No domhelp	76,534,352	399,217	0.0052
With domhelp	2,174,214	77,840	0.0358
Single	72,365,824	414,189	0.0057
Duplex	2,581,274	108,129	0.0419
Apartment/Condo	3,409,318	140,863	0.0413
Industrial/Agricultural Building	335,851	44,832	0.1335
Others	16,297	5,321	0.3265
With Educ	73,269,656	397,556	0.0054
No Educ	5,438,905	137,494	0.0253
Not Elem	15,585,754	180,704	0.0116
With Elem	63,122,808	388,354	0.0062
No Coed	45,207,660	373,362	0.0083
With Coed	33,500,904	325,202	0.0097
With Spouse	67,840,832	392,296	0.0058
No Spouse	10,867,731	144,857	0.0133

the Akaike Information Criterion (AIC) is used which is given by

$$AIC = \frac{-2L + 2\dot{p}}{\tilde{N}}$$

where L is the overall loglikelihood, $\dot{p}$ is the number of covariates in the model (including intercept) and $\tilde{N}$ is the number of cells in the contingency table. Notice that using six variables for the loglinear model gives the best fit. Therefore, using six variables for the ESPREE model is acceptable for generation of the updated small area estimates given the limitations of computer memory capacity mentioned in Section 7.3.1. Incorporating more variables in the model is also a way of minimizing bias

Table 7.6: Some of the models fitted with the corresponding AIC

Number of variables	Variable(s)	AIC
1	wall type	9,915.73
2	wall type and urbanity	16,301.76
4	wall type, urbanity, head_male and all_hsed	4,090.71
6	wall type, roof type, urbanity, head_male, all_hsed and dom_help	748.40

in the estimation procedure given that all the variables included in the model are relevant. Note that although only the margins are changed, the overall model also includes interactions of all order via the census (or pseudo-census) data.

The GLSM method discussed in Section 4.6 was illustrated by Zhang and Chambers (2004) using only a single auxiliary variable and involves estimating an additional parameter in order to reduce bias. Incorporating six variables in the GLSM model is not computationally feasible. However, this is not a problem with the ESPREE method as has been pointed out in Chapter 4, since additional parameter estimation *per se* is not then required.

7.4 Variance Estimation

The details of the theory of variance estimation used in this research are given in Chapter 6. As mentioned, there are two sources of variation for the updated small area estimates - the survey data and the pseudo-census data. Hence, the variance of the updated small area estimates under the ESPREE method takes into consideration the variability from the survey margins and the pseudo-census data, which could be viewed as the sum of the two variances (survey margins variance and pseudo-census variance).

As stated in Section 7.2.1, there was a change in the definition of the primary sampling unit in the 2003 survey from barangay to an area (contiguous barangays) with at least 500 households. However, the difference between the number of PSUs (2826) and barangays (2836) in the original survey is negligible. For simplicity, the barangays are used as the primary sampling units, in this way the survey design fits the balanced repeated replicates (BRR) design so that variance estimation for the survey data is straightforward as presented in Section 6.3.2. Since the survey data has a large number of strata, partial balancing was used.

The variance from the set of pseudo-census data is computed from the set of bootstrap estimates from the small area estimation project based on 2000 survey and census data mentioned in Section 7.2.2, (see also Haslett and Jones, 2005). Computation of the variance from the bootstrap estimates follows the method described in Section 6.3.2 and details of the bootstrapping procedure employed are given in Section 2.3.3. The two variance estimates are then added to get an estimate of the variance of the updated small area estimates.

7.5 Illustration of ESPREE Modelling Procedure

In this Section we illustrate the different steps of the ESPREE modelling procedure presented in Section 5.4 in order to give an overview of the actual steps undertaken to generate the necessary estimates and to show a sample of cell counts both in the pseudo-census and pseudo-counts data sets:

1) The pseudo-census data is generated by employing a modified ELL method using the census and survey data gathered in the year 2000. Here is a sample of the data set for a few cells and 5 replicates:

mcode	Yb	wall	urb	hd_male	all_hsed	roof	dom_hlp	f1	f2	f3	f4	f5
12801	0	1	0	0	0	1	0	15	16	15	6	9
12801	0	1	0	0	0	1	1	0	0	0	0	0
12801	0	1	0	0	0	2	0	0	4	4	0	1
12801	0	1	0	0	0	2	1	0	0	0	0	0
12801	0	1	0	0	0	3	0	0	0	0	0	0
12801	0	1	0	0	0	3	1	0	0	0	0	0
12801	0	1	0	0	0	4	0	0	0	0	0	0
12801	0	1	0	0	0	4	1	0	0	0	0	0

2)-3) Using replicates of the survey margins in the period t_1, replicates of pseudo-counts were generated to facilitate scaling of the pseudo-census counts from step 1 to the appropriate margins by fitting the loglinear model in the next step. The table below shows an example of a set of pseudo-counts with 5 replicates:

mcode	Yb	wall	urb	hd_male	all_hsed	roof	dom_hlp	PS1	PS2	PS3	PS4	PS5
12801	0	1	0	0	0	1	0	199.59	199.63	199.49	199.38	199.50
12801	0	1	0	0	0	1	1	4.72	4.74	4.73	4.730	4.73
12801	0	1	0	0	0	2	0	66.03	66.02	65.98	65.95	65.99
12801	0	1	0	0	0	2	1	1.56	1.57	1.56	1.56	1.56
12801	0	1	0	0	0	3	0	2.44	2.43	2.45	2.43	2.44
12801	0	1	0	0	0	3	1	0.05	0.05	0.05	0.05	0.05
12801	0	1	0	0	0	4	0	38.42	38.40	38.39	38.37	38.38
12801	0	1	0	0	0	4	1	0.91	0.91	0.91	0.91	0.91

4)-5) An example of the loglinear model fitted using the pseudo-census and pseudo-counts cross-classification is given below. An output for step 6 of the ESPREE modelling procedure is given in the next Section.

Variable	Coef.	SE	P-value
Yb	-0.0972	0.0003	0.0000
wall_strong	0.2859	0.0003	0.0000
wall_light	0.4056	0.0004	0.0000
wall_salvaged	0.2819	0.0012	0.0000
roof_strong	-0.9840	0.0004	0.0000
roof_light	-1.0001	0.0005	0.0000
roof_salvaged	-0.3668	0.0014	0.0000
urb	-0.2996	0.0003	0.0000
head_male	-0.4360	0.0003	0.0000
all_hsed	0.0845	0.0003	0.0000
dom_help	0.7578	0.0008	0.0000
_cons	1.0257	0.0005	0.0000

7.6 Intercensal Small Area Estimates of Poverty Incidence

In this Section the intercensal estimates of poverty incidence and the corresponding estimated standard error (SE) and coefficient of variation (CV) of the estimates generated from the ESPREE method using the loglinear model with six auxiliary variables as shown in Section 7.3.3 are presented. The results from the ELL updating method are also presented in order to compare the quality of the estimates generated from the two small area updating methods. The two methods are also compared with the survey-based (FIES) estimates at the provincial and regional levels.

7.6.1 Municipal Level Estimates

A summary of the municipal level (small area) estimates of poverty incidence for both ESPREE and ELL updating methods and their standard errors and CVs are presented in Table 7.7. The mean of the poverty incidence computed from the ESPREE method is higher than the one generated from the ELL updating method. This is further supported by the quantile-quantile plot in Figure 7.1 which shows that the ELL updating method tends to generate lower values of poverty incidence estimates than the ESPREE method in most of the municipalities or small areas.

Table 7.7: Summary of municipal level estimates via ESPREE and ELL

	ESPREE			ELL		
	Incidence	SE	CV	Incidence	SE	CV
Mean	0.4200	0.0418	0.1177	0.3755	0.0413	0.1371
Std. Dev.	0.1713	0.0144	0.0620	0.1843	0.0194	0.0892
Min	0.0204	0.0038	0.0358	0.0114	0.0044	0.0140
Max	0.8937	0.1725	0.5787	0.9746	0.1812	0.8600

The average estimated standard errors of poverty incidence estimates from the ESPREE and the ELL updating methods are similar. However, the average CVs computed from the two methods indicate that the ESPREE method generates more precise estimates than the ELL-based method. The ESPREE-based municipal level estimates of poverty incidence are presented in Figure 7.2.

7.6.2 Provincial Level Estimates

The National Statistical Coordination Board (NSCB) in the Philippines is generating the survey-based provincial level (the third administrative level in the country) estimates of poverty measures. Users of this information are however cautioned on the precision of the estimates since some of the estimates have rather high estimated CV's. Presented in Table 7.8 is a comparison of the estimates generated from FIES (using the survey based estimation procedure in STATA), ESPREE and ELL updating methods. We note that the survey-based estimates generated differ slightly from the official estimates released by the NSCB since the survey-based estimates generated here are based on the combined FIES/LFS data and the PSGC codes used are for the

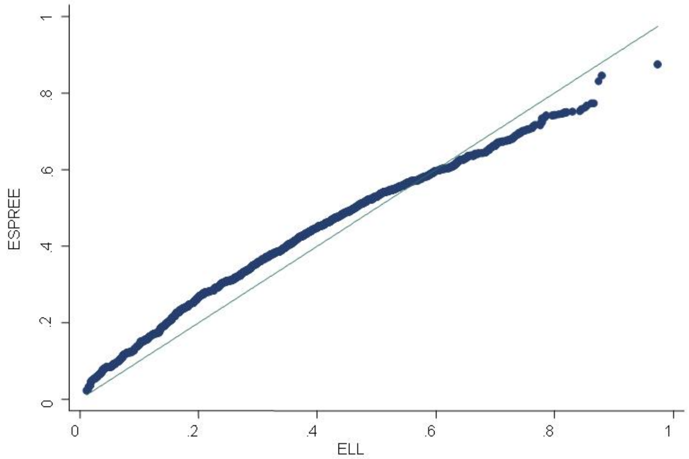


Figure 7.1: Quantile-Quantile plot of ESPREE and ELL updated estimates

year 2000. It appears that the average poverty incidence estimate for the ESPREE and survey-based estimates are close to each other while the ELL-based estimate is lower. This is seen more clearly in the quantile-quantile plot of the ESPREE (Figure 7.3) and ELL (Figure 7.4) estimates versus the survey-based poverty incidence estimates for all the provinces.

Table 7.8: Summary of provincial level estimates via ESPREE and ELL

	Survey-based			ELL			ESPREE		
	Incidence	SE	CV	Incidence	SE	CV	Incidence	SE	CV
Mean	0.3708	0.0417	0.1241	0.3316	0.0181	0.0660	0.3677	0.0173	0.0535
Std. Dev.	0.1526	0.0237	0.0655	0.1490	0.0081	0.0398	0.1440	0.0059	0.0241
Min	0.0530	0.0098	0.0523	0.0302	0.0055	0.0198	0.0457	0.0058	0.0261
Max	0.6851	0.1792	0.4875	0.6804	0.0413	0.2318	0.6408	0.0312	0.1473

It can be observed from Table 7.8 that ESPREE generated the lowest estimate of average estimated standard errors and coefficient of variation but close to the estimates generated from the ELL updating method. It is also clear that the two (ESPREE and ELL updating) methods generated much lower estimated SE and CV than the survey-based estimates. The estimated CV of the survey-based estimates is averaging

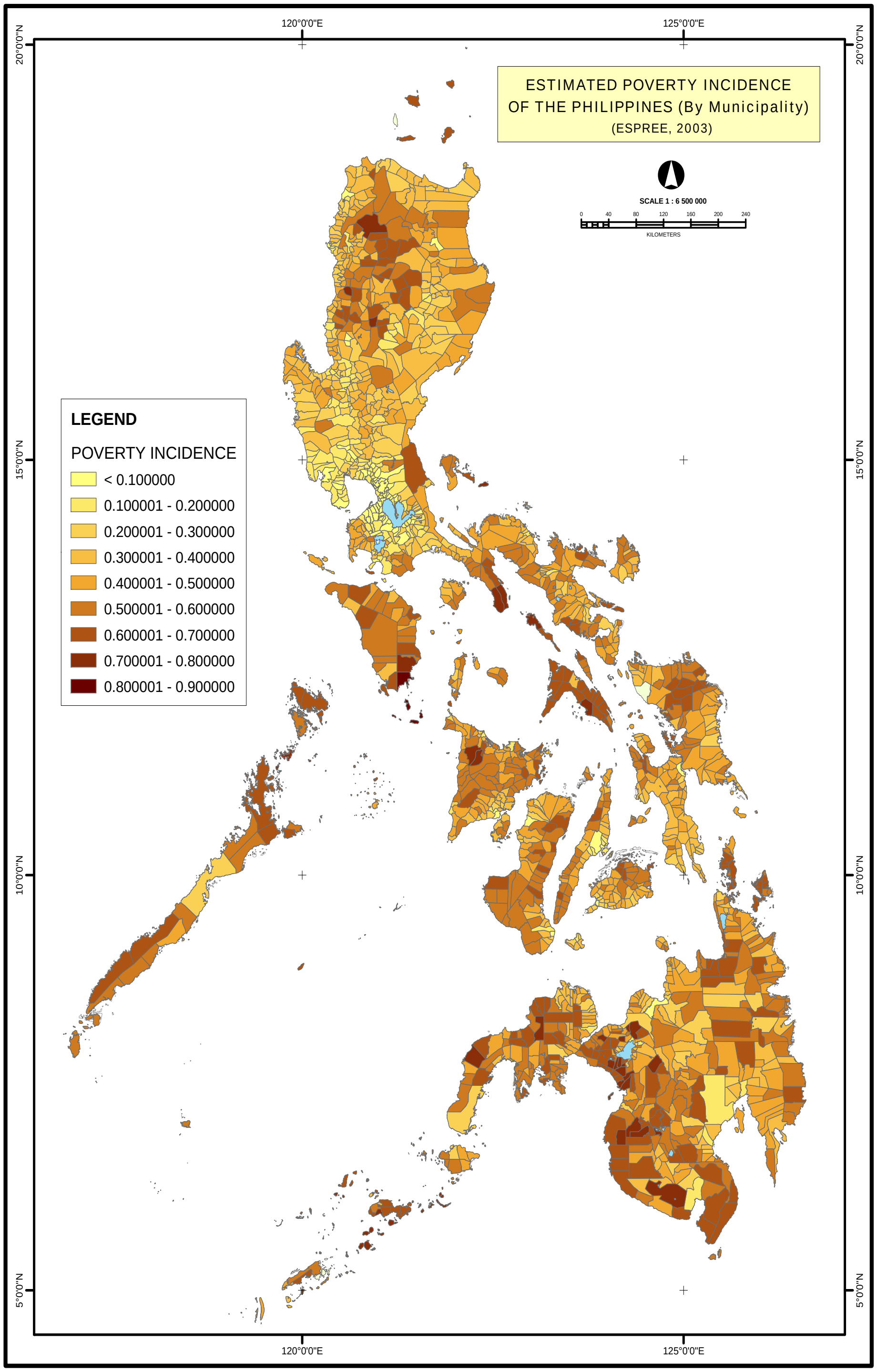


Figure 7.2: 2003 Municipal Level poverty incidence estimates

over 10% while the ESPREE and ELL estimates are averaging around 4% only. The large values of estimated CVs from the survey-based estimates are due to the sample sizes at the provincial level which are too small for accurate estimation.

7.6.3 Regional Level Estimates

The intercensal small area estimates of poverty incidence via ESPREE were also accumulated to generate estimates for the regional level. These regional level estimates were compared with the estimates from the ELL updating method and the survey-based (combined FIES/LFS) estimates (Table 7.9). The differences between the survey-based and ELL as well as between survey-based and ESPREE are summarized by Z scores similar to the one presented in Section 7.3.1 which represents the standardized distance between the two sets of estimates. The Z-scores for the ESPREE estimates are computed as follows:

$$Z = \frac{\text{ESPREE estimate} - \text{FIES estimate}}{\sqrt{(\text{ESPREE standard error})^2 + (\text{FIES standard error})^2}}$$

It is noticeable from Table 7.9 that some regional level estimates from both the ESPREE and ELL updating methods are more than two standard errors away from

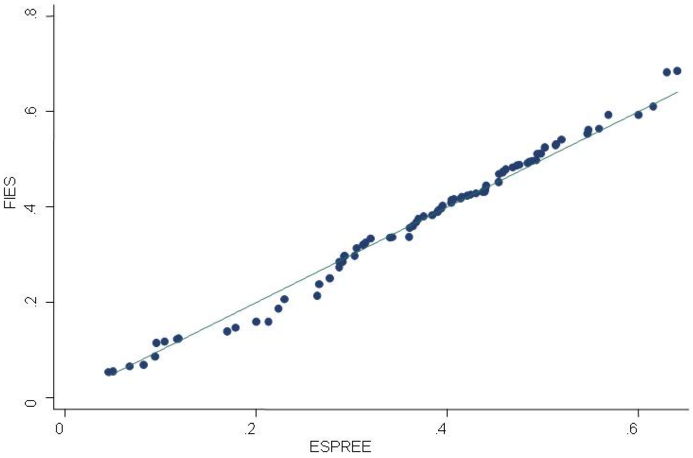


Figure 7.3: Quantile-Quantile plot of ESPREE and FIES provincial level estimates

the corresponding survey-based estimates. This is more common with the ELL updating estimates. In addition, the average of the absolute values of the Z-scores is higher for the ELL updating method, which means that in general the ESPREE method generates regional level poverty incidence estimates closer to the survey-based.

Considering the estimated standard errors computed for the different methods, the ESPREE method tends to have lower estimated standard error compared to the survey-based except for two regions (Region II and Region XI). The ELL updating method tends to have the lowest estimated standard error among the three methods but only conditional on its updating model being correct. To further examine the regional level estimates, quantile-quantile plots were generated as shown in Figures 7.5 and 7.6. It can be observed that the ELL updating poverty incidence estimates tend to be lower than the survey-based estimates.

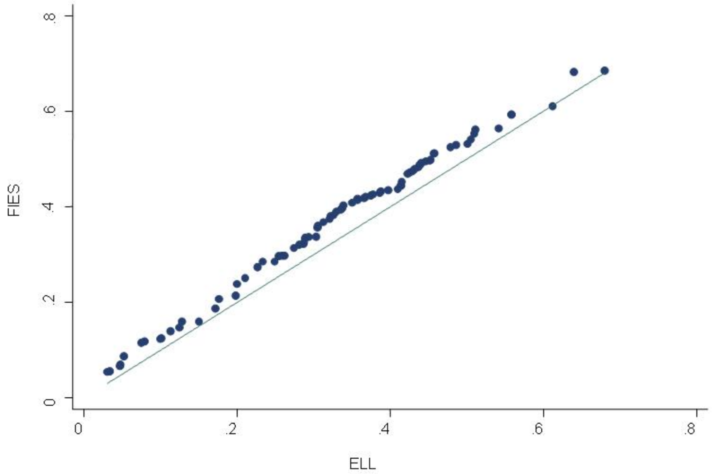


Figure 7.4: Quantile-Quantile plot of ELL and FIES provincial level estimates

Table 7.9: Comparison of regional level poverty incidence estimates

Region	Survey-based		ELL			ESPREE		
	Incidence	SE	Incidence	SE	Z	Incidence	SE	Z
REGION I	0.3030	0.0170	0.2579	0.0144	-2.0260	0.2963	0.0126	0.3156
REGION II	0.2430	0.0134	0.2639	0.0125	1.1402	0.3226	0.0147	-3.9984
REGION III	0.1720	0.0105	0.1386	0.0073	-2.6040	0.1854	0.0073	-1.0472
REGION IV*	0.2443	0.0089				0.2433	0.0067	0.0958
REGION V	0.4845	0.0152	0.3899	0.0119	-4.9032	0.4533	0.0139	1.5145
REGION VI	0.3894	0.0154	0.3243	0.0099	-3.5593	0.3978	0.0102	-0.4563
REGION VII	0.2778	0.0149	0.2717	0.0101	-0.3419	0.3481	0.0128	-3.5841
REGION VIII	0.4303	0.0179	0.4199	0.0125	-0.4760	0.4133	0.017	0.6899
REGION IX	0.4958	0.0205	0.4631	0.0154	-1.2743	0.4320	0.0197	2.2458
REGION X	0.4137	0.0231	0.4212	0.0148	0.2730	0.3369	0.0153	2.7733
REGION XI	0.3490	0.0141	0.3191	0.0155	-1.4303	0.3295	0.0146	0.9651
REGION XII	0.4379	0.0291	0.3600	0.0165	-2.3296	0.4731	0.0137	-1.0948
NCR	0.0697	0.0063	0.0388	0.0053	-3.7406	0.0579	0.0044	1.5211
CAR	0.3290	0.0199	0.271	0.0133	-2.4231	0.343	0.0174	-0.5316
ARMM	0.5520	0.0278	0.4601	0.0274	-2.3535	0.5947	0.0182	-1.2853
REGION XVI	0.5300	0.0200	0.5244	0.0134	-0.2329	0.4712	0.0170	2.2422

*Using the year 2000 Philippine standard geographic codes

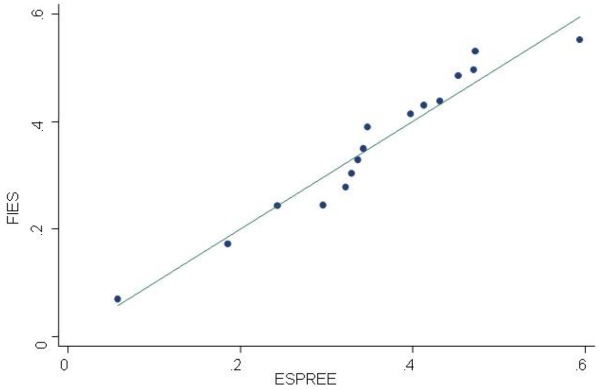


Figure 7.5: Quantile-Quantile plot of FIES and ESPREE regional level estimates

7.7 Discussion

Two intercensal updating methods for the generation of the updated small area estimates of poverty incidence are compared: our proposed ESPREE method, and the

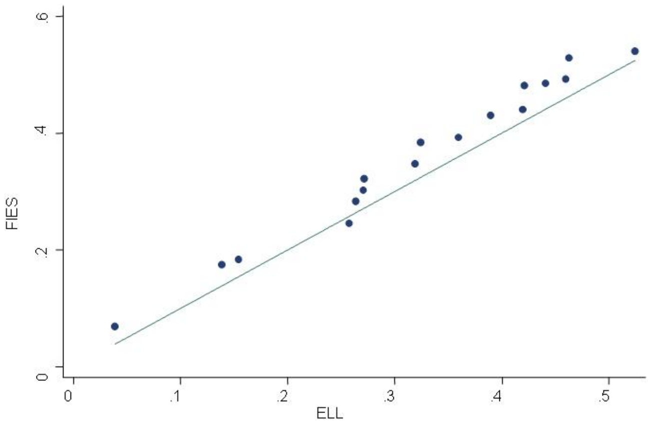


Figure 7.6: Quantile-Quantile plot of FIES and ELL regional level estimates

ELL updating method currently implemented by the World Bank in collaboration with national statistical agencies in the Philippines and Vietnam. For smaller tables another method that could be employed for generating updated small area estimates is the GLSM proposed by Zhang and Chambers (2004). The estimates from the three methods (ESPREE, ELL updating and GLSM), cannot be directly compared because GLSM requires that small area level counts or population be known or accurately estimated from a particular source (e.g., survey) which is not the case for the Philippines. Moreover, fitting a GLSM using a similar set of explanatory variables used for the ESPREE model could lead to intractable parameter inestimability given the sparse survey as emphasized in Section 4.6.

The ESPREE method accounts for structural change from the census year to the most recent period when the survey data is gathered. This is one of the limitations of the ELL updating method as it only uses variables deemed to be time invariant by definition. Thus far there is no appropriate test or method to establish time invariance for the ELL updating method. Most of the auxiliary variables deemed time invariant as shown in Chapter 3 are either barangay or municipal means derived from the census data, only a few household characteristics were included. The choice of the auxiliary variables (aside from other technical issues of the ELL method pointed out in Chapter 2) explains the quality of the poverty incidence estimates generated. Since the auxiliary variables are mostly from the census period, there is not much new or updated information that has been incorporated to the estimation process. Hence, the estimates derived from the ELL updating method are generally lower than for the ESPREE method and the survey-based estimates at the provincial and regional levels.

The ESPREE poverty incidence estimates at the provincial and regional levels are evidently closer to the survey-based estimates. Thus, we can claim that although the ESPREE method may have some limitations, it is not biased and performs better than the ELL updating method. At present, the ESPREE method is only using a model with six auxiliary variables; nevertheless it is able to incorporate the new information from the most recent survey in an optimal manner. Moreover, the inclusion of six auxiliary variables in the ESPREE model is rather more than can be incorporated using the GLSM model since fitting a GLSM or GLSMM model for six variables is computationally infeasible. On balance, it appears that the best method available for updating poverty estimates given new survey but not census data seems to be ESPREE.

Chapter 8

Validation Study

8.1 Introduction

Based on the comparison made in Chapter 7 between the updated small area estimates generated from the ESPREE and ELL updating method, it is clear that the ELL based estimates are biased and hence, substantial differences are observed between the two sets of estimates. The real test of the quality of the estimates however is how well these estimates reflect the actual poverty situation on the ground. Following the analyses, provinces and some municipalities in a selected region (Region I) in the Philippines were visited to conduct validation exercises in order to have a qualitative assessment of the actual performance of the two methods. These validation visits were funded by the New Zealand Postgraduate Study Abroad Awards (NZPSAA) a New Zealand Government scholarship, administered by Education New Zealand. Acceptability and consistency of the estimates were assessed by comparing the estimates with the expert opinion of key informants and their perception on available poverty related indicators at the small area or municipal level. These validation activities are adopted from the validation exercises conducted for the results (small area estimates of poverty measures) of the collaborative poverty mapping project of the World Bank (WB) and National Statistical Coordination Board (NSCB) in the Philippines (NSCB, 2005).

The design of the validation study is discussed in Section 8.2 which includes the mechanics of the validation exercises (Section 8.2.1), the areas covered (Section 8.2.2) and the validation exercise participants (Section 8.2.3). This is followed by the presentation of the results of the validation exercises (Section 8.3) from the different provinces of Region I, starting with the province of Ilocos Norte (Section 8.3.1), followed by the province of Ilocos Sur (Section 8.3.2), province of La Union (Section 8.3.3) and Pangasinan (Section 8.3.4). This Chapter ends with a discussion of the significant insights gained from the exercise and some recommendations (Section 8.4).

8.2 Validation Exercise Design

8.2.1 Mechanics of the Validation Exercise

The validation exercise was carried out by having a one-on-one interview with each of the identified participants, described in detail in Section 8.2.3, using a validation form or questionnaire presented in Appendix G. The questionnaire is an adaptation of the validation form used in the World Bank and NSCB poverty mapping project (NSCB, 2005) which contains poverty related indicators that were included in the set of auxiliary variables used in formulating the ESPREE model, other correlates of poverty and indicators of the Millennium Development Goals (MDGs). Although the questions are structured in a manner that participants were supposed to answer in terms of a score (a number out of 10), some participants were not confident in stating a number and preferred to rank the municipalities. The responses were therefore summarized in terms of mean ranks. The lower the rank the lower the incidence of poverty and the better the situation in a municipality in terms of the indicators.

One of the limitations of the method is that questions on the indicators were answered by the participants based on their perception of the present situation in the municipalities of their province while the estimates they are being compared to are for the year 2003. A question comparing the present with the poverty situation five years ago is included in order to gather some idea of the change or progress in the area. The data gathered on the comparison are included in the tables presented for the results per province in Section 8.3, the column called “Compare 5yrs ago”. There is also an issue about differences in the way poverty is perceived and defined since the ELL and ESPREE-based estimates are based on economic measures only.

Based on the validation form mentioned above, there are two sets of municipal rankings gathered from the participants - (1) the indicator-based and (2) the overall level of poverty assessment. The two sets of municipal ranks were then compared with the ranking of the updated small area estimates generated from the ESPREE and ELL updating method. Discussions were made as to which of the two methods is perceived to generate estimates reflecting the real poverty situation in the municipalities and possible reasons for discrepancies. Participants provided information as to the situation in their localities in terms of thriving industries, livelihood, educational opportunities, among others.

8.2.2 Areas Covered

The validation activity was conducted in the Ilocos Region (Figure 8.1) of the Philippines. Ilocos is located on the northwestern coast of Luzon island, bounded on the east by the Cordillera Administrative Region (CAR) and on the west by the South China Sea. This region includes four provinces: Ilocos Norte, Ilocos Sur, La Union and Pangasinan. In the year 2000, the region's total population was about 4 million, and the province of Pangasinan has the largest population which composes about 58% of its total population. Despite the generally rough terrain of the region, it has very good agricultural land suitable for cultivating crops such as tobacco, rice and various fruits and vegetables. At present, the region is the Philippines' leading producer of tobacco and mangoes for export. This region is also famous for tourism as it houses various Spanish heritage churches and one of UNESCO's World Heritage cities. Since most municipalities are located along the coastal areas fishing and salt making are some of the major sources of income of the residents in the area.

It can be observed from Figure 8.1 that the ESPREE-based municipal level poverty incidence estimates are generally lower in coastal municipalities than those in the mountainous areas. Mountainous areas are usually inaccessible as there are no sealed roads yet which is one of the important infrastructure requirements that would help improve and develop the area. These municipalities are mostly occupied by different aboriginal groups, hence there are some complications in developing these areas as the government also aims to preserve and protect cultural minorities.

One of the advantages of doing the validation study in this region is that it has comparatively stable administrative boundaries at the small area (municipality) level and provinces have not moved from one region to another in the last five years. Moreover, its Regional Development Council (RDC) has recently responded to the call of the Philippine government for improvement of the implementation of poverty alleviation programs in the country. The RDC has created a masterlist of municipalities in the four provinces which are now the beneficiaries for the various poverty alleviation

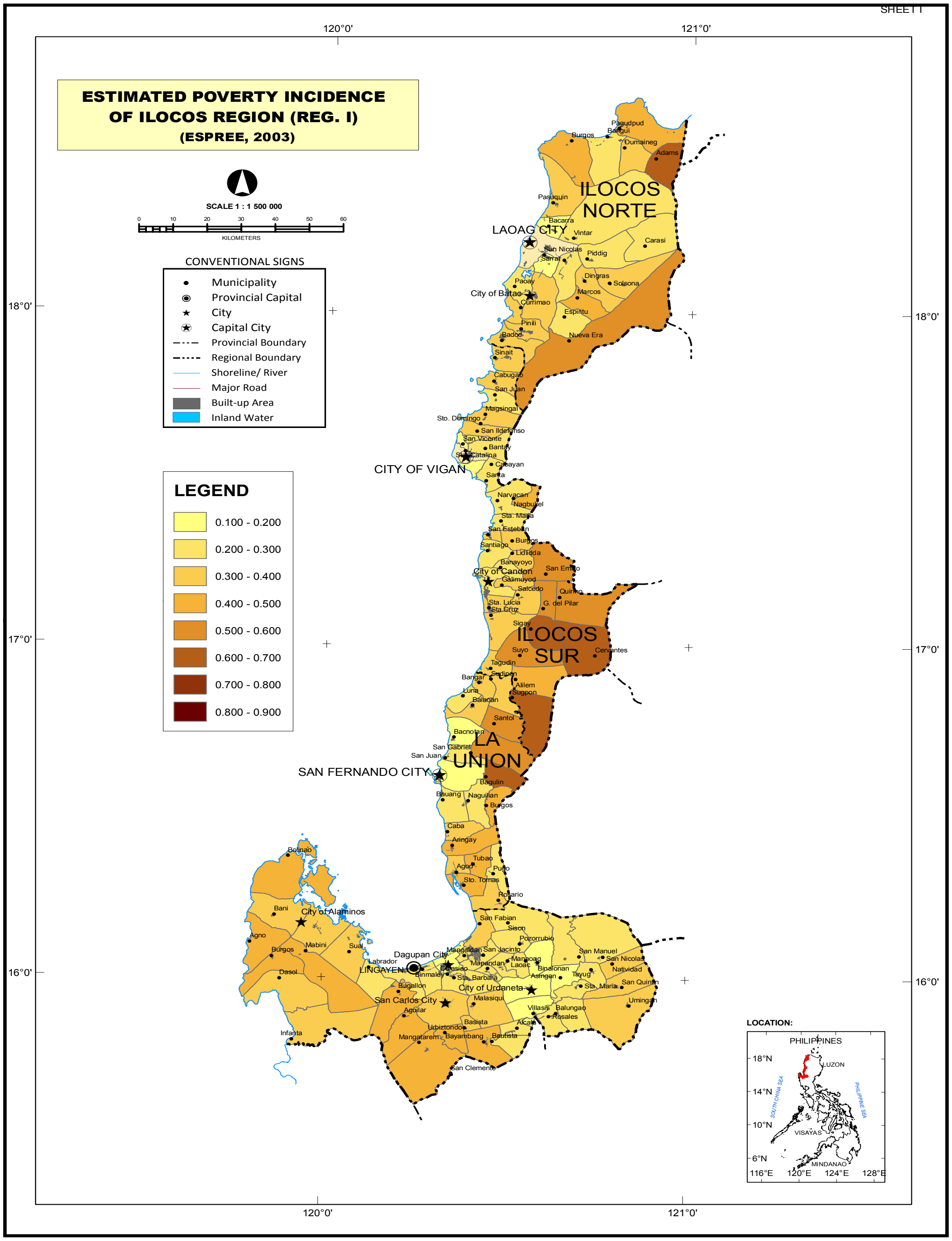


Figure 8.1: Ilocos region with ESPREE municipal level poverty incidence estimates

programs in the region (RDC-I, 2008).

The RDC masterlist was created by using the results of various poverty mapping and small area estimation projects conducted - (1) the small area estimation project carried out by the World Bank in collaboration with the NSCB (NSCB, 2005); (2) the recently implemented intercensal updating of small area estimates project, also by the World Bank and the NSCB (NSCB, 2009); (3) the areas identified as beneficiaries of the Kapit-bisig Laban sa Kahirapan (KALAHI) project (Balisacan et al. (2002), Balisacan and Edillon (2003)) , the Philippine government's poverty alleviation program which started in 2001 ; and (4) the poverty mapping project conducted by the NSCB Regional office to identify the poor areas in the provinces of Region I through the use of social and economic indicators (NSCBR-I, 2000). The municipalities considered poor in any of the four methods were grouped in three priority groups for each province. Those municipalities considered poor in all the methods are listed as the first priority, while those found in two of the methods are included in the list for second priority and those municipalities considered poor in at least one of the methods are listed as third priority.

8.2.3 Validation Exercise Participants

The participants of the validation exercises were composed of representatives from local government units such as the Municipal and Provincial Planning and Development Office, Provincial Social Welfare and Development Office, Provincial Health Office, City Planning and Development Office, offices of National Statistics Office and National Police and. A total of thirty participants from the four provinces were interviewed: seven from Ilocos Norte, nine from Ilocos Sur, six from La Union and eight from Pangasinan.

8.3 Validation Exercises Results

As stated in Section 8.2.1 mean ranks were computed from the validation form (indicator-based and overall assessment of the participants) and compared with the ranking of the updated small area estimates generated from the ESPREE and ELL updating methods. We note that the small area estimates, both from the ESPREE and ELL methods, do not have definitive rankings since some of the estimated standard errors are quite large. However we cannot infer the variability in the ranks from the estimated standard errors of the estimates since these estimates are correlated, and the correlations are not available for the ELL updating method. We note too that large estimated standard errors are more common in the estimates generated from the ELL updating method, reflecting the lack of complexity in the models incorporating only those variables considered to be time-invariant. The overall average estimated standard error is higher for the estimates from the ELL updating method.

Rank correlations of the participants' assessment and the ESPREE and ELL estimates are presented in Table 8.1. The ranking generated from the ESPREE method tends to be in agreement with at least one (indicator-based or overall level of poverty) of the participants' ranking in all the provinces. In addition, among those provinces such that both the ESPREE-based and ELL-based ranks are significantly correlated with the participants' assessment, the ESPREE-based ranking tend to have a higher correlation coefficient estimate, signifying that the participants assessment generally agree with the estimates generated from the ESPREE method more than the estimates generated from the ELL updating method. This could also mean that the indicators or variables used by the participants in coming up with the ranking of the municipalities were considered as predictors (although not as the predicted variable) in the ESPREE model but not in the ELL model. The ESPREE model which can be considered as a "global model", i.e., one model for the whole country, included variables such as education, housing quality, urbanity and presence of household help. The ELL model on the other hand is a region specific model (one model for each region), that has considered presence of a street pattern, number of hotels and similar establishments and education as explanatory variables (NSCB, 2009).

Table 8.1: Rank correlation between participants assessment and the small area estimates (ESPREE and ELL)

	Ilocos Norte		Ilocos Sur		La Union		Pangasinan	
	Rs	p-value	Rs	pvalue	Rs	pvalue	Rs	pvalue
Indicator-based vs ESPREE	0.629	0.001	0.1614	0.3619	0.744	0.000	0.587	0.000
Overall rank vs ESPREE	0.610	0.002	0.3702	0.0312	0.837	0.000	0.062	0.677
Indicator-based vs ELL	0.476	0.022	0.2046	0.2457	0.599	0.005	0.491	0.000
Overall rank vs ELL	0.320	0.136	0.3106	0.0738	0.711	0.000	0.073	0.624

For some of the municipalities in the four provinces the estimates generated from the ESPREE and the ELL updating methods are in agreement. However, as stated earlier, there are municipalities where the estimates are conspicuously opposing. More specific discussion of the ranking discrepancies of the small area estimates generated from the ESPREE and ELL updating methods and the participants assessment are presented in the Sections that follow.

8.3.1 Province of Ilocos Norte

The validation exercises started off in Ilocos Norte, the northernmost part of the region. In this province, the perception of the participants tends to agree more with the ESPREE-based estimates than the ELL-based updating estimates as shown in Table 8.1. The participants' indicator-based ranking and the overall level of poverty ranking from the participants are significantly correlated with the ESPREE-based ranking. The ELL-based ranking is significantly correlated only with the indicator-based ranking but its estimated rank correlation coefficient is lower than the estimate for the ESPREE-based ranking. The higher estimated rank correlation coefficient for ESPREE-based ranking could be due to some of the indicators included in the validation form that were also incorporated in the ESPREE model but not in the ELL model. For example, housing quality, which is not included in the ELL model, is considered in the ESPREE model.

As shown in Table 8.2, there are two municipalities with estimates from ESPREE and ELL updating that are really contradictory, namely Bacarra and Pagudpud. The two sets (indicator-based and overall level of poverty) of rank based on the participants' assessment are both closer to the ESPREE-based rank than the ELL-based rank. The municipality of Bacarra is quite a controversial municipality, as this is included in the top priority list for poverty alleviation projects and one of the bases of selection is the results of the ELL updating method. This municipality is famous in the province for it has the highest number of overseas workers and houses in these areas are in general of good quality as compared to other municipalities. Most participants agree that houses in the province are still considered as a status symbol. Better houses would indicate the owner has the luxury of spending money on more expensive housing materials. The kind of houses in the province is captured in the ESPREE model as housing quality is one of the auxiliary variables used. The participants believe that Bacarra should not really be included in the list of the top priority group for poverty alleviation projects although there could also be some barangays that need assistance within this municipality.

The municipality of Pagudpud is also included in the list of priority municipalities for poverty alleviation. Under the ELL method, this municipality is considered to be one of those with lower poverty incidence in other words a more affluent municipality. However, as pointed out by key informants, the municipality of Pagudpud can only be considered as an " average municipality", i.e., if the municipalities in the province of Ilocos Norte will be grouped into three, one being the group of more affluent municipalities, two for average and three for poor municipalities; Pagudpud should belong to the second group. In terms of agricultural productivity, this municipality has a very small land area for agriculture. Fishing and tourism are its two major livelihoods as it is located on the coastal area of the province. However, this municipality is quite far from the city center and hence, less accessible. In terms of the different indicators considered in the validation form, this municipality may be better off than other municipalities but there are not so many business establishments and tourist facilities in the area for it to be considered a more progressive municipality.

Examining the the fourth and the the seventh column of Table 8.2, we can observe that the municipalities have either improve or maintained their poverty situation as perceived by the participants. These values are averages of the rates (1=improved, 2=maintained, 3= worsened) given by the participants.

8.3.2 Province of Ilocos Sur

For the province of Ilocos Sur, the ELL based ranking is not significantly correlated to the two sets of ranking based on the participants' perception as shown in Table 8.1. The ESPREE-based ranking on the other hand is moderately correlated with (one of the two sets) the participants' assessment of the overall level of poverty in the municipalities. Hence, we can say that in this province, the participants' perception is again in agreement with the ESPREE-based estimates more than the ELL-based

Table 8.2: Validation result for the province of Ilocos Norte

	Based on Participants' Evaluation						Small Area Estimates of Poverty Incidence			
	Different Indicators			Overall Level of Poverty						
	Mean Rank	Overall Rank	Compare 5yrs ago	Mean Rank	Overall Rank	Compare 5yrs ago	ESPREE estimate	Rank	ELL estimate	Rank
ADAMS	22.8	23	1.1	21.8	23	1.3	0.63	23	0.48	21
BACARRA	3.5	3	1.3	5.1	5	1.0	0.21	3	0.33	14
BADOC	12.3	10	1.3	7.8	7	1.0	0.37	16	0.33	13
BANGUI	13.0	13	1.3	13.4	14	1.5	0.25	5	0.23	4
BATAC	2.0	2	1.4	1.4	1	2.0	0.24	4	0.25	5
BURGOS	15.5	17	1.2	15.5	18	1.3	0.33	12	0.27	6
CARASI	19.5	21	1.3	21.3	21	1.7	0.48	21	0.46	20
CURRIMAO	11.5	8	1.3	13.2	13	1.7	0.28	7	0.20	2
DINGRAS	6.0	5	1.2	5.6	6	1.3	0.35	15	0.35	17
DUMALNEG	20.8	22	1.3	21.5	22	1.3	0.30	9	0.31	11
ESPIRITU (Banna)	13.3	14	1.6	15.3	16	2.0	0.38	17	0.30	9
LAOAG CITY	1.0	1	1.4	3.1	2	2.3	0.09	1	0.17	1
MARCOS	17.0	19	1.2	17.1	19	1.3	0.44	19	0.48	22
NUEVA ERA	16.0	18	1.4	15.5	17	2.0	0.56	22	0.55	23
PAGUDPUD	13.5	15	1.2	14.0	15	1.3	0.45	20	0.29	8
PAOAY	12.3	11	1.2	9.7	8	1.7	0.25	6	0.28	7
PASUQUIN	7.5	6	1.3	9.9	9	2.0	0.34	13	0.34	15
PIDDIG	14.8	16	1.2	12.9	12	1.3	0.32	11	0.35	16
PINILI	18.3	20	1.2	18.5	20	1.0	0.38	18	0.38	19
SAN NICOLAS	3.8	4	1.2	4.9	4	1.0	0.13	2	0.20	3
SARRAT	11.8	9	1.3	11.7	11	1.7	0.30	8	0.30	10
SOLSONA	12.5	12	1.4	10.9	10	1.7	0.34	14	0.32	12
VINTAR	7.8	7	1.1	3.6	3	1.0	0.31	10	0.37	18

estimates. For some municipalities however, especially those with higher incidence of poverty, e.g., Sugpon and Sigay, as shown in Table 8.3, the ranking generated from the ELL updating method is close to the ESPREE method and hence they are both in agreement with the participants' perception.

Under the ESPREE method, the city of Vigan (capital of the province) and the municipality of Santa Catalina are the two areas with the lowest incidence of poverty. One the other hand, the lowest poverty municipalities are Banayoyo and Santa Maria under the ELL method. Based on the discussion with the participants, they believe that the ESPREE method is providing more realistic estimates than the ELL method. They believe that Vigan city could possibly be one of those areas with the lowest incidence of poverty, primarily because it is the capital of the province where infrastructure, facilities and services are better than any other municipalities in the province and secondly, because it is one of UNESCO's World Heritage cities and hence, hundreds of tourists are coming into the city every day adding to the earnings of various business establishments in the area.

As to the municipality of Santa Catalina, participants believe that this municipality could be better off than Banayoyo and Santa Maria and could possibly be ranked close to Vigan city since this municipality is similar to the municipality of Bacarra in the province of Ilocos Norte as described in the previous Section. It may not be the general perception that this municipality has low incidence of poverty (as compared to Candon City) but it is the place where most of the overseas workers are living and where houses are generally of better quality compared to other municipalities. In addition, in terms of peace and order, this municipality is one of those considered to be the most peaceful (lowest crime rate) municipalities in the region. In terms of the other indicators, Santa Catalina seemed to have poor performance (rank 17 overall). According to the participants, one possible reason could be that some people in this municipality might have become more dependent on remittances from abroad that could have affected their level of productivity. As observed by the participants there was a drop in enrolment in some of the primary and secondary schools in this area.

There are seven municipalities in this province where the ESPREE-based rank and the ELL-based rank are largely different - the ranks differ by at least ten. Aside from the

Table 8.3: Validation result for the province of Ilocos Sur

	Based on Participants' Evaluation						Small Area Estimates of Poverty Incidence			
	Different Indicators			Overall Level of Poverty						
	Mean Rank	Overall Rank	Compare 5yrs ago	Mean Rank	Overall Rank	Compare 5yrs ago	ESPREE estimate	Rank	ELL estimate	Rank
ALILEM	10.7	21	1.6	23	23	1.0	0.48	27	0.45	31
BANAYOYO	10.9	22	1.6	27	28	2.0	0.30	12	0.11	1
BANTAY	7.6	10	1.7	8.5	9	2.0	0.25	5	0.25	20
BURGOS	11.6	25	1.1	23	24	1.3	0.38	20	0.17	10
CABUGAO	6.8	8	1.6	4.5	3	1.0	0.39	22	0.26	22
CANDON	4.1	1	1.6	1.5	1	1.0	0.29	11	0.16	8
CAOAYAN	11.8	26	1.7	20.5	18	1.3	0.18	3	0.15	7
CERVANTES	9.5	15	1.4	13.5	13	1.3	0.65	32	0.56	33
GALIMUYOD	10.0	18	1.4	22.5	20	2.0	0.37	19	0.27	23
G. DEL PILAR	12.5	27	1.7	28.5	30	2.0	0.56	28	0.34	26
LIDLIDDA	8.5	12	1.4	17	16	2.0	0.36	18	0.13	3
MAGSINGAL	10.1	19	1.6	22	19	1.7	0.34	17	0.28	24
NAGBUKEL	12.8	28	1.6	29	31	2.0	0.48	26	0.38	29
NARVACAN	6.4	6	1.6	6	5	2.0	0.32	14	0.23	17
QUIRINO (ANGKAKI)	11.0	23	1.0	31	33	1.3	0.56	29	0.36	27
SALCEDO (BAUGEN)	7.1	9	1.4	14.5	15	1.3	0.41	23	0.24	19
SAN EMILIO	10.2	20	1.4	25	26	1.3	0.58	30	0.41	30
SAN ESTEBAN	11.0	23	1.4	23	25	2.0	0.32	13	0.14	5
SAN ILDEFONSO	9.7	16	1.6	22.5	21	1.3	0.29	10	0.17	11
SAN JUAN (LAPOG)	7.7	11	1.0	18	17	1.3	0.28	7	0.31	25
SAN VICENTE	10.9	22	1.6	22.5	22	2.0	0.18	4	0.24	18
SANTA	11.2	24	1.4	27.5	29	1.7	0.26	6	0.15	6
SANTA CATALINA	9.8	17	1.9	5	4	1.7	0.05	1	0.19	13
SANTA CRUZ	5.6	3	1.4	14	14	2.0	0.42	25	0.20	16
SANTA LUCIA	8.8	13	1.4	13	12	2.0	0.38	21	0.20	14
SANTA MARIA	7.7	11	1.0	7	6	1.3	0.29	8	0.12	2
SANTIAGO	13.1	29	1.3	25.5	27	2.3	0.29	9	0.20	15
SANTO DOMINGO	9.0	14	1.7	11	11	2.0	0.33	15	0.18	12
SIGAY	13.5	30	1.4	30.5	32	2.0	0.71	34	0.37	28
SINAIT	6.2	5	1.1	8	7	1.3	0.34	16	0.17	9
SUGPON	16.2	31	1.4	31	34	2.0	0.68	33	0.71	34
SUYO	6.5	7	1.7	10	10	2.0	0.58	31	0.47	32
TAGUDIN	4.8	2	1.6	8	8	1.7	0.41	24	0.25	21
VIGAN (Capital)	5.7	4	1.7	1.5	2	2.0	0.06	2	0.14	4

municipalities of Santa Catalina and Banayoyo that were mentioned above, the other municipalities in which the ESPREE and ELL updating methods have generated opposing ranks are San Vicente, San Juan, Lidlidda, Burgos and Bantay. Among these seven municipalities, it is only in San Vicente that the ELL updating method has generated an estimate that is closer to the participants' perception (overall level of poverty assessment) than that from the ESPREE method. This could be due to some of the variables that participants considered in the ranking of the municipalities that were captured in the model used by the ELL method for the region which includes presence of street patterns i.e. networks of streets of at least three streets or roads (NSCB, 2009). According to the municipal representative, at present, they have two major projects in the municipality - infrastructure (road construction) and scholarship projects for out of school youth.

8.3.3 Province of La Union

The province of La Union is the regional center of the Ilocos region. Regional offices are located in the capital of the province - San Fernando city. In this province, the ranking of the municipalities generated from the ESPREE and the ELL updating methods are both generally well and positively correlated to the ranking based on the participants' perception (indicator-based and the overall level of poverty assessment). Once again higher values of the rank correlation are observed between the participants' perception and the ESPREE estimates, which means that participants tend to agree more with the ESPREE estimates than the ELL estimates of poverty incidence in the municipalities of the province.

Examining the ranks of the estimates generated from the ESPREE and the ELL updating methods, the estimates seemed to be in agreement in most municipalities except for the municipality of Santo Tomas. This municipality is one of those considered to have higher incidence of poverty under the ESPREE method which also agrees with the participants' perception. However, the ELL method generated a contradicting result which ranks the municipality among those with low incidence of poverty. As per participants' opinion, the municipality of Santo Tomas cannot be considered to be one of those with lower incidence of poverty. The services and facilities like banks and hospitals, infrastructure and business establishments available in the municipality are not comparable with those available in more affluent municipalities. This municipality has no hospitals or clinic in the area and has only one rural bank. In addition, it is the smallest municipality in terms of land area and is classified as a fourth class municipality in terms of income classification. Income classification of municipalities is usually from one to five: the lower the classification, the higher the income of the municipality.

8.3.4 Province of Pangasinan

In the province of Pangasinan the ranking of the municipalities generated from the ESPREE and the ELL based updating methods are correlated with the participants' indicator-based ranking but not with the overall level of poverty ranking. Moreover, the correlation coefficient of the indicator-based assessment (of the participants) and the ESPREE based estimates is higher than the coefficient for the ELL method. This implies that the participants' opinion tend to be in agreement with the ESPREE-based ranking more than the ELL-based ranking. As pointed out earlier this could be due to the variables included in the set of indicators used in the validation form. On the other hand, the participants' perception of the overall level of poverty in the municipalities is not significantly correlated with either the ESPREE or ELL -based ranking.

There are three municipalities in the province that have opposing ranks (difference of at least 16) between the ESPREE and the ELL updating methods, namely Santa Maria, Bugallon and Alaminos city. Comparing the ranks with the participants' perception, we noticed that the ESPREE-based ranks are closer to the participants ranking of the municipalities based on the different indicators, while the ELL-based ranks are closer to the participants' ranking based on their perception of the overall level of poverty. As with the other provinces, the primary reason for this situation may be the variables that were included in the ESPREE and ELL models. The ESPREE model is closer to the participants' indicator-based ranking because some of the indicators used in the validation form are included in the ESPREE model. The participants' perception of the overall level of poverty however could be based

Table 8.4: Validation result for the province of La Union

	Based on Participants' Evaluation						Small Area Estimates of Poverty Incidence			
	Different Indicators			Overall Level of Poverty						
	Mean Rank	Overall Rank	Compare 5yrs ago	Mean Rank	Overall Rank	Compare 5yrs ago	ESPREE estimate	Rank	ELL estimate	Rank
AGOO	8.0	4	2.0	2.0	2	1.5	0.32	8	0.26	7
ARINGAY	10.5	7	1.5	11.3	17	1.3	0.44	15	0.30	12
BACNOTAN	3.5	2	1.3	3.3	4	1.2	0.21	2	0.21	5
BAGULIN	10.5	7	1.7	11.3	18	1.3	0.68	20	0.86	20
BALAOAN	7.5	4	2.2	3.8	5	1.6	0.34	9	0.25	6
BANGAR	14.5	9	1.7	6.3	9	1.3	0.39	12	0.31	13
BAUANG	8.0	4	1.0	1.5	1	1.0	0.23	4	0.20	2
BURGOS	17.5	11	2.0	11.5	19	1.5	0.48	14	0.43	17
CABA	15.5	10	1.8	8.8	13	1.4	0.37	10	0.26	9
LUNA	10.0	6	1.3	7.3	10	1.2	0.30	7	0.30	11
NAGUILIAN	5.5	3	1.0	2.5	3	1.0	0.28	6	0.26	8
PUGO	8.5	5	1.7	6.0	8	1.3	0.27	5	0.27	10
ROSARIO	8.0	4	1.8	4.5	6	1.4	0.37	11	0.39	16
SAN FERNANDO	2.0	1	1.8	1.5	1	1.4	0.12	1	0.14	1
SAN GABRIEL	12.0	8	1.3	9.0	14	1.2	0.55	18	0.48	18
SAN JUAN	9.0	5	1.7	5.8	7	1.3	0.24	3	0.20	3
SANTO TOMAS	15.0	9	1.2	10.0	15	1.1	0.49	17	0.20	4
SANTOL	18.5	12	1.3	10.3	16	1.2	0.62	19	0.73	19
SUDIPEN	14.5	9	1.3	7.8	11	1.2	0.41	13	0.35	14
TUBAO	11.5	8	1.8	8.3	12	1.4	0.47	16	0.37	15

on other variables that participants considered to be pertinent indicators of poverty that could have been captured by the ELL model.

A very good example of the situation described above is the city of Alaminos. This city's economy partly depends on tourism and as stated earlier, one of the variables included in the ELL model is the number of hotels and similar establishments in the area and as agreed by the participants tourist facilities were indeed one of the indicators they considered in coming up with the overall level of poverty assessment. As to the other indicators which were included in the validation form this city might not be as good as other cities or municipalities, hence, its overall indicator-based rank is closer to the ESPREE based rank.

In addition, this province is nearer to Manila and hence, development in the area is generally faster than in the other provinces in the region. As shown in Table 8.5, from the participants assessment of the comparison of their assessment with the situation five years ago, the majority of the municipalities have improved.

8.4 Discussion and Recommendation

In general the ranking of municipalities based on the participants' perception is in agreement with the ESPREE-based ranking. There are a few municipalities however for which the participants' assessment is closer to the ELL-based ranking, especially the ranking of municipalities based on the overall level of poverty, presumably because the participants considered other variables in coming up with their assessment. Some of these variables deemed by the participants as important indicators were included in the ELL model, e.g., number of hotels and similar facilities and presence of street patterns. We note that one of the major features of the ELL updating method is the use of time-invariant auxiliary variables hence, the data for the auxiliary variables used for generating estimates under the ELL method are from the census period and hence, not reflecting the more recent (survey period) situation leading to the generation of biased estimates of poverty measures. The bias in the estimated poverty incidence however is not necessarily reflected in the ranking, perhaps explaining why some ELL-based ranks were close to the ranking based on participants' assessment,

Table 8.5: Validation result for the province of Pangasinan

	Based on Participants' Evaluation						Small Area Estimates of Poverty Incidence			
	Different Indicators			Overall Level of Poverty						
	Mean Rank	Overall Rank	Compare 5yrs ago	Mean Rank	Overall Rank	Compare 5yrs ago	ESPREE estimate	Rank	ELL estimate	Rank
AGNO	19.8	32	2.2	32	22	1.5	0.42	40	0.46	48
AGUILAR	21.0	35	2.1	29	20	1.7	0.48	46	0.32	38
ALAMINOS CITY	18.7	29	2.5	4	3	1.3	0.31	27	0.21	11
ALCALA	10.8	8	1.9	31.5	21	1.5	0.20	10	0.15	2
ANDA	19.8	32	2.2	34	25	1.3	0.40	38	0.35	41
ASINGAN	9.6	6	1.5	23.5	16	2	0.16	4	0.17	7
BALUNGAO	24.9	39	2.4	43	33	1.6	0.26	21	0.19	10
BANI	18.7	29	2.2	23.5	16	1.1	0.38	34	0.44	46
BASISTA	12.2	11	1.8	38.5	29	1	0.27	23	0.26	24
BAUTISTA	15.1	17	1.8	43	33	2	0.24	15	0.21	13
BAYAMBANG	13.3	13	1.6	9	6	1.5	0.44	43	0.30	34
BINALONAN	7.0	3	1.8	18.5	13	2	0.17	5	0.17	5
BINMALEY	10.8	8	1.8	16	12	1.5	0.27	22	0.21	15
BOLINAO	18.8	30	2.3	11.5	9	1.5	0.50	47	0.46	47
BUGALLON	18.9	31	2.1	20.5	15	1.4	0.44	42	0.28	26
BURGOS	18.0	26	2.1	40	31	1.5	0.46	45	0.36	42
CALASIAO	8.0	4	1.7	11	8	1.5	0.20	12	0.24	21
DAGUPAN CITY	6.7	2	1.8	1.5	1	1.6	0.11	1	0.11	1
DASOL	21.1	36	2.4	36	26	1.3	0.39	36	0.29	33
INFANTA	18.6	28	1.9	33	24	1.5	0.32	29	0.30	35
LABRADOR	16.7	23	1.9	43	33	1.5	0.25	17	0.17	4
LAOAC	25.7	40	2.0	43.5	34	1	0.26	19	0.28	31
LINGAYEN (Capital)	9.1	5	1.8	11.5	9	1.2	0.20	11	0.23	20
MABINI	22.9	37	2.3	32	22	1.2	0.45	44	0.37	44
MALASIQUI	13.0	12	1.6	7	5	1.3	0.39	37	0.34	40
MANAOAG	15.5	20	1.9	20.5	15	1.5	0.24	16	0.21	14
MANGALDAN	8.0	4	1.8	7	5	1.5	0.16	3	0.22	17
MANGATAREM	15.4	19	1.8	9.5	7	1.5	0.43	41	0.31	37
MAPANDAN	14.1	16	2.0	37	27	2	0.18	7	0.21	12
NATIVIDAD	23.7	38	2.0	41.5	32	1	0.32	28	0.28	30
POZZORUBIO	13.9	15	1.9	11.5	9	1	0.30	25	0.25	22
ROSALES	15.2	18	2.1	19.5	14	2	0.26	20	0.23	18
SAN CARLOS CITY	9.6	6	2.3	3	2	1.4	0.39	35	0.28	29
SAN FABIAN	15.9	22	2.1	14	10	1	0.33	31	0.29	32
SAN JACINTO	18.0	26	1.9	27.5	19	1	0.30	26	0.28	27
SAN MANUEL	10.8	8	1.6	19.5	14	1	0.30	24	0.27	25
SAN NICOLAS	20.4	34	1.5	25.5	18	1	0.32	30	0.32	39
SAN QUINTIN	17.4	24	2.2	32.5	23	1.5	0.36	32	0.28	28
SANTA BARBARA	11.7	9	1.9	11.5	9	2	0.23	13	0.23	19
SANTA MARIA	13.6	14	1.4	39.5	30	1.5	0.24	14	0.30	36
SANTO TOMAS	20.2	33	2.3	46	35	2	0.19	8	0.17	6
SISON	11.9	10	2.3	24.5	17	2	0.18	6	0.21	16
SUAL	15.9	21	1.5	5	4	1	0.38	33	0.37	45
TAYUG	18.4	27	1.9	25.5	18	2	0.25	18	0.18	8
UMINGAN	15.4	19	1.9	15.5	11	1.5	0.41	39	0.26	23
URBIZTONDO	17.6	25	1.8	37.5	28	1.5	0.50	48	0.36	43
URDANETA CITY	6.1	1	1.3	1.5	1	1	0.13	2	0.16	3
VILLASIS	10.7	7	2.3	16	12	2	0.19	9	0.18	9

especially some of the participants' overall level of poverty ranks. We note however that among the significant rank correlations, the correlation coefficients for the ESPREE-based ranking with any of the two participants-based rankings were higher than those for the ELL-based ranking.

It can be observed from the tables of ranks from the four provinces presented in the previous four Sections that there are also some discrepancies in the ranking of the municipalities based on the indicators and the ranking for overall level of poverty. As pointed out by the participants, they have considered other variables or indicators in coming up with the ranking of the municipalities for overall level of poverty. These indicators were not necessarily listed in the set of indicators included in the validation form. Some variables considered were:

- location and accessibility of municipalities and other infrastructure available in the area like seaport and airport
- presence of cultural minorities
- availability of various livelihood in the area related to agriculture, fishery and tourism, business and financial establishments (e.g., hotels and banks) and health facilities (e.g., hospitals and clinics)

The overall level of poverty ranking from the participants' perspective is related to the Participatory Poverty Index (PPI), a composite measure of poverty level which incorporates the views of poor households on what they consider to be critical indicators of poverty (Xiaoyun and Remenyi, 2008).

Moreover, participants are trying to emphasize that having family members working as overseas workers has a great impact on the socio-economic situation of families in their localities. Various studies have been conducted in the Philippines linking poverty and overseas remittances (see for example, Martinez and Yang (2005), and Yang and Choi (2007)). Remittances to the Philippines from overseas in 2007 were around 1.2 billion USD per month (FORBES, 2008). Data on remittances could also be included in the ESPREE model, however, this variable is not yet available in the census data at hand. These various correlates of poverty that were suggested by the validation exercise participants for inclusion in the ESPREE model are usually available at the municipal and/or barangay (cluster) level, but at present not all municipalities in the country are as diligent as some of the municipalities in Region I who collect annual data on the said variables. It would be helpful if the administrative data on these pertinent indicators would be updated in all the municipalities of the country.

In addition, the comparison made of the participants' perception of the present poverty condition with the situation five years ago shows that across provinces the majority perceived that either the poverty situation of their municipalities have improved or maintained. This could be considered as an indication that the government's poverty alleviation programs are working well in most of these areas. The survey-based estimate of poverty incidence for the whole region showed an improvement from about 36% in the year 2000 to 30% in 2003. The ESPREE-based estimate is around 30% as well, while the ELL-based estimate is way much lower at about 26%. The ESPREE and survey-based estimates are more reflective of having some municipalities either maintaining or improving their poverty situation than the ELL-based estimate which is about 10% improvement from the year 2000.

Overall, the validation exercises did not only play an integral part in the assessment of the quality of the estimates generated from the two intercensal updating procedures but allowed us to gather more insights into other factors deemed as important poverty indicators by key informants residing in the area. The indicators they suggested, when incorporated into the estimation process, could make a great improvement to the existing methods of estimating poverty measures which are generally economic-based, e.g., income and consumption-based poverty measures. It should be noted however that the indicators considered by the participants led to their overall assessment of poverty and their ranking of the municipalities. This is different to both the ESPREE and ELL updating methods which generate an estimate of an economic variable (income/consumption) based on a number of variables and use the predicted income/consumption to generate estimates of poverty measures and hence the poverty level ranking of municipalities.

Chapter 9

Concluding Remarks

9.1 Introduction

General results and conclusions of this research are summarized in this Chapter. Recommendations for future research activities aimed at improving methods of generating small area estimates of poverty measures in Third World countries are also included.

9.2 Summary of Results and Conclusions

This research primarily aims to develop an updating method for small area estimates of poverty measures in Third World countries. The problem of updating or generating small area estimates of poverty measures during non-census years is an offshoot of the small area estimation procedures for poverty measures in Third World countries that use census data, such as the ELL method, which requires a survey and a census assumed to have been conducted at the same time period. In Third World countries however a census is usually conducted only once every 10 years while a national survey is conducted once every three years. The survey-plus-census based methods such as the ELL method in its original form therefore cannot be used for generating small area estimates when we have a new survey but no new census conducted at the same time period as the national survey. In addition, the ELL method which is the aid-industry standard has theoretical issues that have to be addressed and improved given its role in aid allocation and poverty monitoring in Third World countries. As emphasized in the introductory Chapter, this dissertation covers two main topics: 1) theoretical issues of the ELL method and 2) the development of an updating method for poverty measures in Third World countries.

The ELL method appears to have been developed separately from the "mainstream" or "standard" small area estimation methods. This method involves fitting a model for household level income/expenditure. However the authors did not clearly relate this model to the existing mainstream small area models and they developed a parameter estimation or survey fitting procedure that is different from the procedures used for mainstream small area estimation. This thesis has put the ELL income/expenditure model in the context of mainstream small area estimation models as described by Rao (2003), focusing on linear mixed models, the class of models in which the ELL income/consumption model belongs. The mainstream small area models related to the ELL model are therefore reviewed in Chapter 1. Moreover, since the proposed method for updating poverty measures involves the use of categorical variables (poor and non-poor), a more general framework for small area estimation is therefore established in Chapter 1 namely the generalized linear mixed model (GLMM) framework.

Under the linear mixed models framework, mainstream small area models account for area level effects and models are fitted to the variable of interest. The ELL method on the other hand fits a model to the household level income or expenditure and not directly to a particular poverty measure (the variable of interest). In ELL, the predicted household level income/expenditure are then transformed to generate estimates of poverty measures. In addition, ELL's household level income/expenditure model only incorporates cluster (PSU in survey design) level effects. Clusters are usually smaller than the local level or small area of interest. In the Philippines for example, the small area of interest (municipality) could be composed of at least 4 clusters (barangays). Estimates of poverty measures for small areas are generated by aggregating the household level predicted values. The ELL method could therefore be improved by incorporating small area level effects to the household level income/expenditure model.

The ELL method uses a "weighted generalized least squares" to fit the income or expenditure model. This method incorporates survey weights in the estimation procedure similar to the generalized regression estimation (Lohr, 1999) procedure in an attempt to formulate an estimation procedure that accounts for the heteroscedastic variance as well as the sampling weights. However, the manner in which the survey weights enter the calculation of parameter is rather simplistic given that the variance-covariance matrix used by ELL is not a diagonal matrix. The ELL estimation method ends up with an asymmetric variance-covariance matrix for the estimated regression parameters. In addition, the ELL method has its own method of computing the variance components that incorporates the sampling weights in computing the cluster level variance and usually uses a heteroscedastic model for the household level variance but not at higher levels in the hierarchy of random effects. Given the theoretical limitations of the ELL method, model fitting procedures usually employed under mainstream small area estimation method such as the pseudo-EBLUP, IWEE and GSR should be used to replace the model fitting stage of the ELL method. These methods do not have the theoretical shortcomings of the ELL method.

The application of the different methods to actual survey data set from the Philippines showed that despite the theoretical limitations of the ELL model fitting procedure, it has generated regression parameter estimates that are in general similar to the ones generated from the other methods. The difference lies in the cluster level variance component estimate, which was shown to have the greatest influence in the estimated variance of the prediction error of the regression model. The ELL method has its own method of generating estimated variance components and results showed that its estimated cluster level variance component is the smallest. As noted earlier, this means that (especially in conjunction with the lack of a small area error component) estimated standard errors from the ELL method are not statistically conservative.

As noted above, the ELL method, in its original form and assumptions, cannot be used to generate small area estimates of poverty measures during non-census or intercensal years and hence cannot be used for updating poverty estimates between censuses. Some recently proposed methods are based on the ELL method, either extensions or modifications of the original method, which we call here “ELL-based updating” methods. Given the theoretical limitations of the ELL method pointed out above, it is therefore necessary to develop an alternative updating method. Our proposed updating method is called ESPREE as it is an extension of the SPREE method that can be used for fitting high-dimensional tables and in which the census data is assumed stochastic rather than fixed. This method is compared with other methods that could also be used for updating such as the classical SPREE, one of its extensions developed by Zhang and Chambers (2004) called GLSM or GLSMM, and the ELL-based updating methods.

Classical SPREE can be used to generate updated small area estimates but the ESPREE method has the advantage of allowing for a stochastic association structure which may, at least in part, reduce the bias coming from the assumption of a fixed association structure or having the data from the census assumed to be measured without error that is necessary under SPREE. Another method that has a great potential for reducing the bias in the original or classical SPREE is the use of GLSMs. These GLSMs aims to reduce bias by estimating a proportionality coefficient that accounts for changes in the association structure from the census period to the survey period. However, formulation of GLSMs can be very complicated for high-dimensional tables and relatively sparse survey data set. Fitting of GLSMs for such tables is also computationally difficult if not prohibitive. Moreover, GLSMs require that the small area counts be accurately estimated during the survey period which is not usually the case for data sets available in Third World countries for poverty estimation.

Even putting aside the inherent limitations of the ELL model and its survey fitting procedure, ELL-based updating methods also have some data requirements that are rarely satisfied in most Third World countries. Recently implemented ELL-based updating methods require either panel survey data or time-invariant variables for cross-sectional survey data. At present, the statistical procedure for properly assessing time invariance for the auxiliary variables does not exist. The implementation in the Philippines (Lanjouw and van der Wiede, 2006) of the ELL-based updating method using time-invariant variables had to rely on the researchers' personal judgement to decide whether a variable is time-invariant or not. There is also the complication that even if these time-invariant variables were known, using only the set of time-invariant variables limits the number of candidate models. The ESPREE method on the other hand does not have such stringent data requirements on the auxiliary variables and the available survey data. Moreover, ESPREE allows for "time-varying" variables which are expected to be more useful for explaining changes in the variable interest.

Comparison of the updated small area estimates of poverty measures in the Philippines generated from the ELL-based updating method using time invariant variables and the ESPREE method showed that the ELL-based updated estimates are biased and less precise. The validation study conducted in one of the regions in the Philippines showed that the key informants' assessment of the poverty situation in their area is generally in agreement with or closer to the ESPREE-based than to the ELL-based updated estimates. Hence, when comparing the two methods, the ESPREE method appears to generate estimates of better quality (unbiased and more precise) that are better able to reflect the real poverty situation on the ground.

Based on the comparison of the different methods (ESPREE, ELL-based updating, GLSM and classical SPREE) that can be used for updating small area estimates of poverty measures, in terms of assumptions, data requirements, and applicability to poverty estimation in Third World countries, the ESPREE method appears to be the best available method so far.

9.3 Recommendations and Future Directions

Similar to any other research endeavors, the end of this research has opened up new questions and new research problems to be answered. Given the importance of small area estimates of poverty measures in aid allocation and in monitoring progress towards the Millennium Development Goals it is very important that the method used for generating such estimates is theoretically sound mathematically and statistically and in agreement with acceptable standards for small area estimation. The investigation conducted is only part of a more extensive examination needed if the ELL method can justifiably maintain its role as the official method for generating small area estimates of poverty measures in Third World countries. For example, one further possible area of investigation is comparison of the different survey fitting procedures under a linear mixed model for household level income or expenditure which accounts for area level effects in addition to cluster and household levels.

Under the ELL method, the formulation of the linear mixed model for household level income or expenditure is based on the assumption that small areas are independent; this assumption is however not necessarily true for different small areas (e.g., Philippines' municipalities). Correlation between adjacent small areas needs further examination. If there is sufficient evidence to show that the poverty situation in adjacent small areas (e.g., municipalities) tends to be more similar than those far apart or significant correlation can be established, then linear mixed models with area level random effects may not be sufficient and random effects at an even higher level may be warranted. Other models could be considered that can account for the spatial correlation, e.g. conditional autoregressive (CAR) models, although care is needed with these models since fitting extra correlated area based random effects can complicate model diagnostics and hide problem with underlying models.

In this thesis, the ESPREE method has been shown to be better than the ELL-based updating method, theoretically and in application to real data. However, there are still several avenues for improvement. The current model used for ESPREE could be further improved by including area level effects to account for the correlation between households within small areas (i.e., fitting a GLMM). Another way of improving the model would be to incorporate spatial variation in a similar manner as mentioned above for the ELL household level income or expenditure model, i.e. fitting a GLM with a CAR model to account for the spatial correlation between small areas. One of the recently proposed models that accounts for spatial heterogeneity called geographically weighted regression (GWR) could also be considered. In addition, research should be conducted on effective ways to combine ESPREE based updated estimates with estimates from other data sources or estimation techniques that could lead to improvement of small area updated estimates of poverty measures.

Further investigation is also needed on various diagnostic or evaluation methods for assessing competing small area updating methods or small area models. For example, while it is already apparent that the ESPREE method has various advantages over the ELL-based updating methods, those arguments and evidence could be investigated further by developing other statistical diagnostic procedures primarily designed for small area updating models for poverty measures in Third World countries. Exploration of some of the diagnostics employed by Brown et al. (2001) and more recently by Inglese et al. (2008) should be conducted. Moreover, validation studies comparing key informants' assessment of the poverty situation in their area and the updated estimates should also be conducted in more regions of the country for a more extensive assessment of the acceptability of the estimates and their ability to reflect the real poverty situation on the ground.

In Chapter 6, various estimation procedure of the variance for the ESPREE based estimates have been presented. The BRR method combined with the bootstrap method have been used in the application of the ESPREE method to the Philippine data. These two methods were used in combination for simplicity of implementation, given that the sets of pseudo-census data were available in the form of bootstrap estimates from a previous poverty mapping project (Haslett and Jones, 2005) and the sampling design and the survey data available conform to what is required for a replication method such as BRR. Comparison of the different procedures needs to be conducted to assess their performance in measuring the variance of the updated small area estimates of poverty measures in Third World countries.

Bibliography

ACTIONAID-International (2006). Participatory poverty assessment, attapeu, lao pdr. A publication of the Mekong Wetlands Biodiversity Conservation and Sustainable Use Program.

ADB (2001). Participatory poverty assessment in cambodia. *http://www.adb.org/Documents/Books/Participatory_Poverty/Participatory_Poverty.pdf*.

ADB (2006). Poverty definition, measurement, and analysis. *http://www.adb.org/Statistics/Poverty/glossary.asp*.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics, Hoboken, New Jersey.

Balisacan, A. M. and Edillon, R. G. (2003). Second poverty mapping and targeting for phases iii and iv of kalahi-cidss: Final report. Unpublished report.

Balisacan, A. M., Edillon, R. G., and Ducanes, G. M. (2002). Poverty mapping and targeting for kalahi-cidss: Final report. Unpublished report.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83:28–36.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Massachusetts.

Blackorby, C. and Donaldson, D. (1980). Ethical indices for the measurement of poverty. *Econometrica*, 48:1053–1060.

Bogue, D. J. (1950). A technique for making extensive postcensal estimates. *The Journal of American Statistical Association (JASA)*, 45(5):149–162.

Brooks, S. P. (1998). Markov chain monte carlo method and its appplication. *The Statistician*, 47:69–100.

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001). Evaluation of small area estimation methods - an application to unemployment estimates from the uk lfs. *Proceedings of Statistics Canada Symposium*.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press, Boca Raton, 3rd edition.

Chakravarty, S. R. (1983). A new index of poverty. *Mathematical Social Sciences*, 6(3):307.

Chambers, R. (2006). What is poverty? who asks? who answers? *Poverty in Focus*, UNDP:December 2006, 3–4.

Cressie, N. (1992). Reml estimation in empirical bayes smoothing of census undercount. *Survey Methodology*, 18:75–94.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Elbers, C., Lanjouw, J., and Lanjouw, P. (2002). Micro-level estimation of poverty and inequality. *Research Working Paper 2911, World Bank, Development Research Group, Washington, D.C.*

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

FORBES (2008). Philippines 2007 overseas workers' remittances at record 14.4 billion dollars. *http://www.forbes.com/feeds/afx/2008/02/15/afx4659876.html*.

Foster, J., Greer, J., and Thorbeck, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52:761–766.

Foster, J. and Shorrocks, A. F. (1991). Subgroup consistent poverty indices. *Econometrica*, 59:687–709.

Frankel, M. R. (1971). Inference from survey samples. Int. Social Res., Univ. Michigan, Ann Arbor.

Fujii, T. (2003). Commune-level estimation of poverty measures and its application in cambodia. Preprint.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9:55–93.

Godambe, V. (1991). *Estimating Functions*. Oxford University Press, Inc., New York.

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

Gonzalez, M. E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section*, pages 33–36.

Gonzalez, M. E. and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73:7–15.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25:489–504.

Haslett, S. (1990). Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables. *Computational Statistics and Data Analysis*, 9:179–195.

Haslett, S., Green, A., and Zingel, C. (1998). Small area estimation given regular updates of census auxiliary variables. Proceedings of the New Techniques and Technologies for Statistics Conference. Sorrento, Italy.

Haslett, S. and Jones, G. (2004). Local estimation of poverty and malnutrition in bangladesh. *Bangladesh Bureau of Statistics and United Nations World Food Programme.*

Haslett, S. and Jones, G. (2005). Local estimation of poverty in the philippines. *Philippine National Statistics Co-ordination Board / World Bank Report, (*http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/ Local_Estimation_of_Poverty_Philippines.pdf*).*

Haslett, S. J., Isidro, M. C., and Jones, G. (2010). Comparison of survey regression techniques in the context of small area estimation of poverty. *To be published in Statistics Canada publication Survey Methodology, Catalogue 12-001-XIE2010002, December 2010, vol. 36 no. 2.*

Haslett, S. J. and Jones, G. (2006). Small area estimation of poverty, caloric intake and malnutrition in nepal. *Published: NepalCentral Bureau of Statistics / World Food Programme, United Nations / World Bank, September 2006, 184pp, ISBN 999337018-5.*

Haslett, S. J., Noble, A., and Zabala, F. (2007). New approaches to small area estimation of unemployment. Project Report for Official Statistics Research Programme - Statistics New Zealand.

Henderson, C. R. (1953). Estimation of variance and variance components. *Biometrics*, 9:226–252.

Hoogeveen, J., Emwanu, T., and Okwi, P. (2003). Updating small area welfare indicators in the absence of a new census. Preprint.

Inglese, F., Russo, A., and Russo, M. (2008). Diagnostics of small area model-based estimators. *http://www.sis-statistica.it/files/pdf/atti/rs08_spontanee_a_2_5.pdf.*

Isaki, C. and Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:89–96.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15(1):1–96.

Jitsuchon, S. and Lanjouw, P. (2005). Updating poverty maps: Emerging experience with available methods and data from thailand. Preprint.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6(3):223–239.

Kakwani, N. (1980). On a class of poverty measures. *Econometrica*, 48(2):438–439.

Kanji, G. K. and Chopra, P. K. (2007). Poverty as a system: Human contestability approach to poverty measurement. *Journal of Applied Statistics*, 34(9):1135–1158.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17:521–526.

Kovar, J. G. (1985). Variance estimation of nonlinear statistics in stratified samples. *Methodology Branch Working Paper*, (87-004E). Statistics Canada.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9:1010–1019.

Lanjouw, P. and van der Wiede, R. (2006). Determining changes in welfare distributions at the micro-level: Updating poverty maps. Powerpoint presentation at the NSCB Workshop for the NSCB/World Bank Intercensal Updating Project.

Larbi, G. A. (1999). The new public management approach and crisis states. UNRISD Discussion paper No. 112.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* Duxbury Press, Brooks/Publishing Company.

Longford, N. T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society Ser. A*, 162:227–245.

Marker, D. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15:1–24.

Martinez, C. and Yang, D. (2005). Remittances and poverty in migrants' home areas: Evidence from the philippines. *SDT - Departamento de Economia, Universidad de Chile*, 257.

Mellor, R. W. (1973). Subsample replication variance estimation. Unpublished PhD Thesis, Harvard University, Cambridge, Massachussets.

Militino, A. F., Ugarte, M. D., Goicoa, T., and Gonzalez-Audicana, M. (2006). Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:450–461.

Minot, N., Baulch, B., and Epprecht, M. (2003). Poverty and inequality in vietnam: Spatial patterns and geographical determinants. *International Food Policy Research Institute, Washington, D.C. and Institute of Development Studies.* University of Sussex.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384.

Neri, L., Ballini, F., and Betti, G. (2005). Poverty and inequality mapping in transition countries. *Statistics in Transition*, 7(1):135–157.

Noble, A. (2003). Small area estimation through glm. Unpublished PhD Thesis, Massey University, New Zealand.

Noble, A., Haslett, S. J., and Arnold, G. (2002). Small area estimation via generalized linear models. *Journal of Official Statistics*, 18:45–60.

NRC (2000). Small-area estimates of school-age children in poverty: Evaluation of current methodology. C. F. Citro and G. Kalton (Eds.), Committee on National Statistics, Washington, DC: National Academy Press.

NSCB (2000). Profile of censuses and surveys. National Statistical Coordination Board, Philippines.

NSCB (2005). Estimation of local poverty in the philippines. *National Statistical Coordination Board, Philippines.*

NSCB (2009). 2003 city and municipal level poverty estimates. *National Statistical Coordination Board, Philippines.*

NSCBR-I (2000). Provincial poverty maps. *http://www.nscb.gov.ph/ru1/poverty_maps.htm.*

Osberg, L. and Xu, K. (2008). How should we measure poverty in a changing world? methodological issues and chinsese case study. *Review of Development Economics*, 12(2):419–441.

Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337.

Pfefferman, D., Moura, F. A., and Silva, P. L. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93:949–959.

Pfefferman, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc B*, 60:23–40.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85:163–171.

Purcell, N. J. (1979). Efficient estimation for small domains: a categorical data analysis approach. Unpublished PhD Thesis. University of Michigan.

Purcell, N. J. and Kish, L. (1979). Estimation for small domains. *Biometrics*, 35:365–384.

Purcell, N. J. and Kish, L. (1980). Postcensal estimates for local areas. *International Statistical Review*, 48:3–18.

Purcell, N. J. and Linacre, S. (1976). Techniques for the estimation of small area characteristics. Unpublished Paper, Australian Bureau of Statistics, Canberra.

Qiao, C. (2006). Small area estimates produced using loglinear models in sas - discussion and implementation. *http://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/ working-papers/wp-01-06-report-loglinear-modelling.doc.*

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. R. Statist. Soc. B*, 11:68–84.

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25:175–186.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley Series in Methodology, New York.

RDC-I (2008). Approving the common masterlist of priority program benefeciaries (cmbpp) for region i poverty reduction program. RDC Resolution no. 57 S. 2008.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–51.

Saei, A., Zhang, L. C., and Chambers, R. (2005). Generalized structure preserving estimation for small areas. *Statistics in Transition*, 7(3):685–696.

Schaible, W. L. (1996). *Use of Small Area Statistics in U.S. Federal Programs.* Springer-Verlag, Inc., New York.

Sen, A. K. (1976). Poverty: An ordinal approach to measurement. *Econometrica*, 44(2):219–231.

Siddhartha, C. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Simonoff, J. S. (2003). *Analyzing Categorical Data.* Springer texts in Statistics, New York.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Survey Data.* John Wiley, Chichester.

Stukel, D. M. and Rao, J. N. K. (1999). Small area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, 78:131–147.

Taylor, M. F. (2001). *British Household Panel Survey user manual*, volume B1. Codebook.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Ann. Math. Statist.*, 29.

UN-website (2009). Millenium development goals. *http://www.un.org/milleniumgoals/bkdg.shtml.*

Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Xiaoyun, L. and Remenyi, J. (2008). Making poverty mapping and monitoring participatory. *Development in Practice*, 18(4-5):599–610.

Yang, D. and Choi, H. (2007). Are remittances insurance? evidence from rainfall shocks in the philippines. *The World Bank Economic Review*, 21(2):219–248.

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Survey Methodology*, 30:431–439.

You, Y. and Rao, J. N. K. (2003). Pseudo hierarchical bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111:197–208.

You, Y., Rao, J. N. K., and Kovacevic, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. *Statistics Canada International Symposium Series-Proceedings.*

Zhang, L. C. and Chambers, R. L. (2004). Small area estimates for cross-classifications. *J. R. Statist. Soc. B*, 66:479–496.

Zhao, Q. (2006). User manual for povmap version 1.1a. *http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf.*

Appendices

Appendix A

Do Files in Stata for Different Survey Fitting Methods

ELL_no_hetero.do

```
// This is the program for fitting the regression model using the ELL method with
random errors assumed to be homoscedastic.

set mem 100m
set matsize 7000
cd “C:\SAE1\Results_ELL”
use “E:\SAE\SURVEY2_all.dta”, replace
/*Re-scale the weights so it would sum up to the sample size*/
egen totwt=total(sswgthh)
gen rwt=sswgthh*(_N/totwt)
egen trwt=total(rwt)
#delimit ;
/*OLS regression using re-scaled survey weights*/
global Xvars “famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong
wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed
all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg
Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52”;
#delimit cr
regress lnincpp $Xvars [pweight=rwt]
global rmseB=e(rmse)
/*Save residuals, coefficients and covariance matrix*/
predict resB, resid
matrix beta=e(b)
matrix VarB=e(V)
/*Calclulate the number of regressors*/
global nx=0
foreach var of varlist $Xvars      {
global nx=$nx+1
}
display $nx

/*Calculate the variance component - cluster level (var-upsilon)*/
by bcode, sort: gen nb=_N
global nb=nb
egen upsilon=mean(resB), by(bcode)
egen wtb=total(rwt), by(bcode)
gen epsilon=resB-upsilon
```

```
egen epsmn=mean(epsilon), by(bcode)
gen df_eps2=(epsilon-epsmn)*(epsilon-epsmn)
gen wb=wtb/trwt
preserve
collapse nb trwt upsilon (sum)sumdeps2=df_eps2 wtb=rwt u=resB, by(bcode)
gen taub2=(1/(nb*(nb-1)))*sumdeps2 /*computation based on SAS prog, but differ from the one in the appendix of the ELL paper*/
gen wb=wtb/trwt
gen nume1=wb*(upsilon*upsilon)
gen denom=wb*(1-wb)
gen nume2=denom*taub2
egen Snume1=total(nume1)
egen Snume2=total(nume2)
egen Sdenom=total(denom)
gen var_ups=(Snume1-Snume2)/Sdenom
egen totres=total(u)/*just to check the sum of resids*/
display totres
egen twb=total(wb) /*just to check the sum of the adjusted weigths(=1)*/
display twb
display var_ups
scalar vU=var_ups[1]
global Nbcode=_N /*number of bcode or clusters*/ restore
gen wtdres=resB*rwt
egen twtdres=total(wtdres)
display twtdres
egen totres1=total(resB)
display totres1
gen var_eps=($rmseB*$rmseB)-vU /*no heteroscedasticity model with location effect*/
/* if no location effect*/
replace epsilon=resB if vU<0
replace var_eps=$rmseB*$rmseB if vU<0 /*no heteroscedasticity model & no location effect*/
scalar vE=var_eps[1]
display vE
/*————————GLS estimation——————————*/
egen bcode1=group(bcode)
by bcode1, sort: gen numbcode=_N
gen const=1
global XvarsC “$Xvars const”
global mb=$nx+1
matrix XWBX=J($mb,$mb,0)
matrix XWBWX=J($mb,$mb,0)
matrix XWBY=J($mb,1,0)
```

```
//have to collapse numbcode by bcode1 and save it and merge it with the data set
or do a preserve and restore command.
forvalues i=1/$Nbcode      {
mkmat $XvarsC if bcode1==‘i’, matrix(X‘i’)
mkmat lnincpp if bcode1==‘i’, matrix(Y‘i’)
mkmat sswgthh if bcode1==‘i’, matrix(W‘i’)
mkmat var_eps if bcode1==‘i’, matrix(varE‘i’)
matrix vE‘i’=diag(varE‘i’)
matrix vU‘i’=vU*J(numbcode1[‘i’],numbcode1[‘i’] ,1)
matrix Wt‘i’=diag(W‘i’)
matrix WB‘i’=Wt‘i’*inv(vE‘i’+vU‘i’)
matrix XWBX=XWBX+(X‘i’)’*WB‘i’*(X‘i’)
matrix XWBY=XWBY+(X‘i’)’*WB‘i’*(Y‘i’)
matrix XWBWX=XWBWX+(X‘i’)’*WB‘i’*(diag(W‘i’))*X‘i’ }
matrix M=inv(XWBX)
matrix Beta=(M*XWBY) matrix Varbeta=M*XWBWX*M
matrix covB=0.5*(Varbeta+Varbeta’)
forvalues i=1/$mb      {
display sqrt(covB[‘i’,‘i’])
}
svmat Beta, name(beta)
preserve
keep beta1
keep if beta1< .
save beta_ELL, replace
restore
svmat covB, name(Var)
preserve
keep Var1-Var31
keep if Var1< .
save covBELL, replace
restore
//generation of the components of the variance of prediction error

//MUNICIPAL LEVEL
sort bcode gen mcode=int(bcode/1000)
sort mcode egen mcode1=group(mcode)
by bcode1, sort: gen numhb=_N
by mcode1, sort: gen numhmun=_N
preserve
collapse bcode1, by(mcode1)
global Nmun=_N
restore
matrix xVxmun=J($Nmun,1,0)
```

```
forvalues i=1/$Nmun {
mkmat $XvarsC if mcode1==‘i’, matrix(Xmun‘i’)
mkmat const if mcode1==‘i’, matrix(consmun‘i’)
matrix xVxmun[‘i’,1]=xVxmun[‘i’,1]+consmun‘i”*Xmun‘i’*covB*Xmun‘i”*consmun‘i’
}
//convert the matrix to single obs and generate the xcovBx data
svmat xVxmun, name(xVxmun)
preserve
keep xVxmun1
keep if xVxmun1< .
gen mcode1=_n
save xVxmunELL, replace
restore
//to generate number of clusters per municipality
preserve
collapse mcode, by(bcode)
by mcode, sort: gen Snclusmun=_N
collapse Snclusmun, by(mcode)
save nclusmunELL, replace
restore
//to generate file containing number of hh
preserve
collapse mcode1 nhh=numhmun, by(mcode)
save mcodenhELL, replace
restore
//PROVINCIAL LEVEL
sort prov
egen prov1=group(prov)
by prov1, sort: gen numhprov=_N
preserve
collapse bcode1, by(prov1)
global Nprov=_N
restore
matrix xVxprov=J($Nprov,1,0)
forvalues i=1/$Nprov {
mkmat $XvarsC if prov1==‘i’, matrix(Xprov‘i’)
mkmat const if prov1==‘i’, matrix(consprov‘i’)
matrix xVxprov[‘i’,1]=xVxprov[‘i’,1]+consprov‘i”*Xprov‘i’*covB*Xprov‘i”*consprov‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxprov, name(xVxprov)
preserve
keep xVxprov1
keep if xVxprov1< .
```

```
gen prov1=_n
save xVxprovELL, replace
restore
//to generate number of clusters per province
preserve
collapse prov, by(bcode)
by prov, sort: gen Snclusprov=_N
collapse Snclusprov, by(prov)
save nclusprovELL, replace
restore
//to generate file containing number of hh
preserve
collapse prov1 nhh=numhprov, by(prov)
save provnhELL, replace
restore
//REGIONAL LEVEL
sort regn egen reg1=group(regn)
by reg1, sort: gen numhreg=_N
preserve
collapse bcode1, by(reg1)
global Nreg=_N
restore
matrix xVxreg=J($Nreg,1,0)
forvalues i=1/$Nreg      {
mkmat $XvarsC if reg1==‘i’, matrix(Xreg‘i’)
mkmat const if reg1==‘i’, matrix(consreg‘i’)
matrix xVxreg[‘i’,1]=xVxreg[‘i’,1]+consreg‘i”*Xreg‘i’*covB*Xreg‘i”*consreg‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxreg, name(xVxreg)
preserve
keep xVxreg1
keep if xVxreg1< .
gen reg1=_n
save xVxregELL, replace
restore
//to generate number of clusters per region
preserve
collapse regn reg1, by(bcode)
by reg1, sort: gen Snclusreg=_N
collapse Snclusreg, by(reg1)
save nclusregELL, replace
restore
//to generate file containing number of hh
```

```
preserve
collapse nhh=numhreg, by(reg1)
save regnhELL, replace
restore
//MUNICIPAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusmunELL”, clear
sort mcode
save nclusmunELL, replace
use “mcodenhELL”
sort mcode
merge mcode using “nclusmunELL”
drop _merge
save mcodenhELL, replace
//combine cluster, num HH and beta effect in one file
use “xVxmunELL”, replace
sort mcode1
save xVxmunELL, replace
use “mcodenhELL”
sort mcode1
merge mcode1 using “xVxmunELL”
gen beta_efctm=xVxmun1/(nhh*nhh)
drop _merge
save xVxmunELL, replace
//PROVINCIAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusprovELL”, clear
sort prov
save nclusprovELL, replace
use “provnhELL”
sort prov
merge prov using “nclusprovELL”
drop _merge
save provnhELL, replace
//combine cluster, num HH and beta effect in one file
use “xVxprovELL”, replace
sort prov1
save xVxprovELL, replace
use “provnhELL”
sort prov1
merge prov1 using “xVxprovELL”
gen beta_efctp=xVxprov1/(nhh*nhh)
drop _merge
save xVxprovELL, replace
```

```
//REGIONAL LEVEL COMBINING FILES
//combine cluster, num HH and beta effect in one file
use “nclusregELL”, clear
sort reg1
save nclusregELL, replace
use “regnhELL”
sort reg1
merge reg1 using “nclusregELL”
drop _merge
save regnhELL, replace
//combine cluster, num HH and beta effect in one file
use “xVxregELL”, replace
sort reg1
save xVxregELL, replace
use “regnhELL”
sort reg1
merge reg1 using “xVxregELL”
gen beta_efctp=xVxreg1/(nhh*nhh)
drop _merge
save xVxregELL, replace
/*to generate the Census clusterand hh effect as well as the file for variance component
estimates*/
use “E:\Size_bgy”, clear
gen nh2=nh*nh
gen nh2vU=(nh2*vU)
egen reg1=group(regn)
//MUNICIPAL LEVEL
preserve
by mcode, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(mcode)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctmunELL, replace
restore
//to generate varcomponents
preserve
use “xVxmunELL”, replace
sort mcode
save xVxmunELL, replace
use “CHHefctmunELL”
sort mcode
merge mcode using “xVxmunELL”
keep if _merge==3
drop _merge
```

```
save varcompsmunELL, replace
restore
//PROVINCIAL LEVEL
preserve
by prov, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(prov)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctprovELL, replace
restore
//to generate varcomponents
preserve
use “xVxprovELL”, replace
sort prov
save xVxprovELL, replace
use “CHHefctprovELL”
sort prov
merge prov using “xVxprovELL”
drop _merge
save varcompsprovELL, replace
restore
//REGIONAL LEVEL
preserve
by reg1, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(reg1)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctregELL, replace
restore
//to generate varcomponents
preserve
use “xVxregELL”, replace
sort reg1
save xVxregELL, replace
use “CHHefctregELL”
sort reg1
merge reg1 using “xVxregELL”
drop _merge
save varcompsregELL, replace
restore
scalar list
```

ELL_w_hetero

```
// This is the program for fitting the regression model using the ELL method with
random errors assumed to be heteroscedastic.
set mem 100m
set matsize 7000
cd “C:\SAE1\Results_ELLH” use “E:\SAE\SURVEY2_all.dta”, replace /*Re-scale
the weights so it would sum up to the sample size*/
egen totwt=total(sswgthh)
gen rwt=sswgthh*(_N/totwt)
egen trwt=total(rwt)
#delimit ;
/*OLS regression using re-scaled survey weights*/ global Xvars “famsize famsizesqc
type_mult per_kids roof_light per_61up roof_strong wall_light wall_salvaged wall_strong
fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed all_coed dom_help head_male no_spouse
hou_9600 hea_rel_mus Per_eng Hou_coelpg Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52”;
#delimit cr
regress lnincpp $Xvars [pweight=rwt]
global rmseB=e(rmse)
/*Save residuals, coefficients and covariance matrix*/
predict resB, resid
matrix beta=e(b)
matrix VarB=e(V)
/*Calclulate the number of regressors*/
global nx=0
foreach var of varlist $Xvars       {
global nx=$nx+1
}
display $nx
/*Calculate the variance component - cluster level (var-upsilon)*/
by bcode, sort: gen nb=_N
global nb=nb
egen upsilon=mean(resB), by(bcode)
egen wtb=total(rwt), by(bcode)
gen epsilon=resB-upsilon
egen epsmn=mean(epsilon), by(bcode)
gen df_eps2=(epsilon-epsmn)*(epsilon-epsmn)
gen wb=wtb/trwt
preserve
collapse nb trwt upsilon (sum)sumdeps2=df_eps2 wtb=rwt u=resB, by(bcode)
gen taub2=(1/(nb*(nb-1)))*sumdeps2 /*computation based on SAS prog, but is different
from the one in the appendix*/
gen wb=wtb/trwt
gen nume1=wb*(upsilon*upsilon)
gen denom=wb*(1-wb)
gen nume2=denom*taub2
egen Snume1=total(nume1)
egen Snume2=total(nume2)
egen Sdenom=total(denom)
gen var_ups=(Snume1-Snume2)/Sdenom
egen totres=total(u) /*just to check the sum of resids*/
display totres
egen twb=total(wb) /*just to check the sum of the adjusted weigths(=1)*/
display twb
display var_ups
scalar vU=var_ups[1]
global Nbcode=_N /*number of bcode or clusters*/
restore
gen wtdres=resB*rwt
egen twtdres=total(wtdres)
display twtdres
egen totres1=total(resB)
display totres1
gen var_eps=(rmseB*rmseB)-vU /*no heteroscedasticity model with location effect*/
/* if no location effect*/
replace epsilon=resB if vU<0
replace var_eps=$rmseB*$rmseB if vU<0 /*no heteroscedasticity model and no location effect*/
/*————-heteroscedasticity modelling————*/
/*computation of var_eps if hetero modelling will be performed*/
gen eps2=epsilon*epsilon su eps2, meanonly
global A=1.05*r(max)
gen lneA=ln(eps2/($A-eps2))
replace lneA=-15 if lneA<-15
//stepwise selection of variables for Zvar
//sw regress lneA $Xvars [pweight=rwt], pe(.05)
global Zvars “all_coed famsize famsizesqc dom_help per_kids per_61up roof_strong wall_light fa_xxl fa_xs all_hsed Per_eng head_male hea_rel_mus Hou_coelpg”
//computed R-square is 0.03
regress lneA $Zvars [pweight=rwt] /* PovMap does not say anything how to choose the Zvar variables*/
predict yhat
rename yhat yhatA
predict resA, resid
matrix alpha=e(b)
matrix VarA=e(V)
global rmseA=e(rmse)
global sigmar2=$rmseA*$rmseA
gen C=exp(yhatA)
```

```
replace var_eps=(($A*C)/(1+C))+($sigmar2/2)*(($A*C)*(1-C)/(1+C)ˆ3)
egen varE=mean(var_eps) scalar vE=varE[1]
display vE
/*——————GLS estimation———————*/
egen bcode1=group(bcode)
by bcode1, sort: gen numbcode=_N
gen const=1
global XvarsC “$Xvars const”
global mb=$nx+1
matrix XWBX=J($mb,$mb,0)
matrix XWBWX=J($mb,$mb,0)
matrix XWBY=J($mb,1,0)
//have to collapse numbcode by bcode1 and save it (like excel file numbcode1) and
paste it on the data set
forvalues i=1/$Nbcode       {
mkmat $XvarsC if bcode1==‘i’, matrix(X‘i’)
mkmat lnincpp if bcode1==‘i’, matrix(Y‘i’)
mkmat sswgthh if bcode1==‘i’, matrix(W‘i’)
mkmat var_eps if bcode1==‘i’, matrix(varE‘i’)
matrix vE‘i’=diag(varE‘i’)
matrix vU‘i’=vU*J(numbcode1[‘i’],numbcode1[‘i’] ,1)
matrix Wt‘i’=diag(W‘i’)
matrix WB‘i’=Wt‘i’*inv(vE‘i’+vU‘i’)
matrix XWBX=XWBX+(X‘i’)’*WB‘i’*(X‘i’)
matrix XWBY=XWBY+(X‘i’)’*WB‘i’*(Y‘i’)
matrix XWBWX=XWBWX+(X‘i’)’*WB‘i’*(diag(W‘i’))*X‘i’ }
matrix M=inv(XWBX)
matrix Beta=(M*XWBY)
matrix Varbeta=M*XWBWX*M
matrix covB=0.5*(Varbeta+Varbeta’)
forvalues i=1/$mb       {
display sqrt(covB[‘i’,‘i’])
}
svmat Beta, name(beta)
preserve
keep beta1
keep if beta1< .
save beta_ELLH, replace
restore
svmat covB, name(Var)
preserve
keep Var1-Var31
keep if Var1< . save covBELLH, replace
restore
```

```
//generation of variance components
//MUNICIPAL LEVEL
sort bcode gen mcode=int(bcode/1000)
sort mcode egen mcode1=group(mcode)
by bcode1, sort: gen numhb=_N
by mcode1, sort: gen numhmun=_N
preserve
collapse bcode1, by(mcode1)
global Nmun=_N
restore
matrix xVxmun=J($Nmun,1,0)
forvalues i=1/$Nmun       {
mkmat $XvarsC if mcode1==‘i’, matrix(Xmun‘i’)
mkmat const if mcode1==‘i’, matrix(consmun‘i’)
matrix xVxmun[‘i’,1]=xVxmun[‘i’,1]+consmun‘i”*Xmun‘i’*covB*Xmun‘i”*consmun‘i’
}
//convert the matrix to single obs and generate the xcovBx data
svmat xVxmun, name(xVxmun)
preserve
keep xVxmun1
keep if xVxmun1< .
gen mcode1=_n
save xVxmunELLH, replace
restore
//to generate number of clusters per municipality
preserve
collapse mcode, by(bcode)
by mcode, sort: gen Snclusmun=_N
collapse Snclusmun, by(mcode)
save nclusmunELLH, replace
restore
//to generate file containing number of hh
preserve
collapse mcode1 nhh=numhmun, by(mcode)
save mcodenhELLH, replace
restore
//PROVINCIAL LEVEL
sort prov egen prov1=group(prov)
by prov1, sort: gen numhprov=_N
preserve
collapse bcode1, by(prov1)
global Nprov=_N
restore
matrix xVxprov=J($Nprov,1,0)
```

```
forvalues i=1/$Nprov      {
mkmat $XvarsC if prov1==‘i’, matrix(Xprov‘i’)
mkmat const if prov1==‘i’, matrix(consprov‘i’)
matrix xVxprov[‘i’,1]=xVxprov[‘i’,1]+consprov‘i”*Xprov‘i’*covB*Xprov‘i”*consprov‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxprov, name(xVxprov)
preserve
keep xVxprov1
keep if xVxprov1< .
gen prov1=_n
save xVxprovELLH, replace
restore
//to generate number of clusters per province
preserve
collapse prov, by(bcode)
by prov, sort: gen Snclusprov=_N
collapse Snclusprov, by(prov)
save nclusprovELLH, replace
restore
//to generate file containing number of hh
preserve
collapse prov1 nhh=numhprov, by(prov)
save provnhELLH, replace
restore
//REGIONAL LEVEL
sort regn egen reg1=group(regn)
by reg1, sort: gen numhreg=_N
preserve
collapse bcode1, by(reg1)
global Nreg=_N
restore
matrix xVxreg=J($Nreg,1,0)
forvalues i=1/$Nreg      {
mkmat $XvarsC if reg1==‘i’, matrix(Xreg‘i’)
mkmat const if reg1==‘i’, matrix(consreg‘i’)
matrix xVxreg[‘i’,1]=xVxreg[‘i’,1]+consreg‘i”*Xreg‘i’*covB*Xreg‘i”*consreg‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxreg, name(xVxreg)
preserve
keep xVxreg1
keep if xVxreg1< .
gen reg1=_n
```

```
save xVxregELLH, replace
restore
//to generate number of clusters per region
preserve
collapse regn reg1, by(bcode)
by reg1, sort: gen Snclusreg=_N
collapse Snclusreg, by(reg1)
save nclusregELLH, replace
restore
//to generate file containing number of hh
preserve
collapse nhh=numhreg, by(reg1)
save regnhELLH, replace
restore
//MUNICIPAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusmunELLH”, clear
sort mcode
save nclusmunELLH, replace
use “mcodenhELLH”
sort mcode
merge mcode using “nclusmunELLH”
drop _merge
save mcodenhELLH, replace
//combine cluster, num HH and beta effect in one file
use “xVxmunELLH”, replace
sort mcode1
save xVxmunELLH, replace
use “mcodenhELLH”
sort mcode1
merge mcode1 using “xVxmunELLH”
gen beta_efctm=xVxmun1/(nhh*nhh)
drop _merge
save xVxmunELLH, replace
//PROVINCIAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusprovELLH”, clear
sort prov
save nclusprovELLH, replace
use “provnhELLH”
sort prov
merge prov using “nclusprovELLH”
drop _merge
save provnhELLH, replace
```

```
//combine cluster, num HH and beta effect in one file
use “xVxprovELLH”, replace
sort prov1
save xVxprovELLH, replace
use “provnhELLH”
sort prov1
merge prov1 using “xVxprovELLH”
gen beta_efctp=xVxprov1/(nhh*nhh)
drop _merge
save xVxprovELLH, replace
//REGIONAL LEVEL COMBINING FILES
//combine cluster, num HH and beta effect in one file
use “nclusregELLH”, clear
sort reg1
save nclusregELLH, replace
use “regnhELLH”
sort reg1
merge reg1 using “nclusregELLH”
drop _merge
save regnhELLH, replace
//combine cluster, num HH and beta effect in one file
use “xVxregELLH”, replace
sort reg1
save xVxregELLH, replace
use “regnhELLH”
sort reg1
merge reg1 using “xVxregELLH”
gen beta_efctp=xVxreg1/(nhh*nhh)
drop _merge save xVxregELLH, replace
/*to generate the Census clusterand hh effect as well as the file for variance component
estimates*/
use “E:\Size_bgy”, clear
gen nh2=nh*nh
gen nh2vU=(nh2*vU)
egen reg1=group(regn)
//MUNICIPAL LEVEL
preserve
by mcode, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(mcode)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctmunELLH, replace
restore
//to generate varcomponents
```

```
preserve
use “xVxmunELLH”, replace
sort mcode
save xVxmunELLH, replace
use “CHHefctmunELLH”
sort mcode
merge mcode using “xVxmunELLH”
keep if _merge==3
drop _merge
save varcompsmunELLH, replace
restore
//PROVINCIAL LEVEL
preserve
by prov, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(prov)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctprovELLH, replace
restore
//to generate varcomponents
preserve
use “xVxprovELLH”, replace
sort prov
save xVxprovELLH, replace
use “CHHefctprovELLH”
sort prov
merge prov using “xVxprovELLH”
drop _merge
save varcompsprovELLH, replace
restore
//REGIONAL LEVEL
preserve
by reg1, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(reg1)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctregELLH, replace
restore
//to generate varcomponents
preserve
use “xVxregELLH”, replace
sort reg1
save xVxregELLH, replace
use “CHHefctregELLH”
```

```
sort reg1
merge reg1 using “xVxregELLH”
drop _merge
save varcompsregELLH, replace
restore
```

pseudo_eblup.do

```
//This program is the implementation of the Pseudo-EBLUP method
clear
set mem 200m
set matsize 5000
cd “C:\SAE1\Results_YR”
use “E:\SAE\SURVEY2_all.dta”, replace
#delimit ;
global Xvars “famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52”;
#delimit cr
/*to generate sigmaE */
global nx=0
foreach var of varlist $Xvars      {
egen ‘var’mn=mean(‘var’), by(bcode)
gen new_‘var’=‘var’-‘var’mn
global nx=$nx+1
}
egen incmean=mean(lnincpp), by(bcode)
gen newinc=lnincpp-incmean
regress newinc new_*
predict e_ij, resid
egen bn=group(bcode)
gen e2=e_ij*e_ij
egen SS1=total(e2)
/*to generate the denominator for SigmaE*/
by bcode, sort: generate bigN=_N
preserve
collapse (mean) bigN $Xvars, by(bcode)
gen id=_n
global num=_N
generate const=1
global XvarsMC “$Xvars const”
foreach var of varlist $XvarsMC quad    {
gen new_‘var’=bigN*‘var’
}
matrix accum B = new_*, noconstant
matrix list B
display $num
restore
gen varE=(1/(_N-$num-($nx+1)+1))*SS1
display varE
display $nx
//to generate var upsilon
gen val=$num-$nx+1
display val
display _N
matrix accum XpX = $Xvars
matrix list XpX
matrix invXX = syminv(XpX)
matrix list invXX
matrix F=invXX*B
matrix list F
matrix T=trace(F)
matrix list T
regress lnincpp $Xvars
predict u_ij, resid
gen uij2=u_ij*u_ij
egen ssuij=total(uij2)
gen varU=(ssuij-((_N-($nx+1))*varE))/(_N-T[1,1])
display varU
display varE
/*to generate beta*/
egen wtsum=sum(sswgthh)
gen nwt=sswgthh/wtsum
egen wijtot=total(sswgthh), by(bcode)
gen wij=sswgthh/wijtot
gen wij2=wij*wij egen dltai2=total(wij2), by(bcode)
gen gmai=varU/(varU+(varE*dltai2))
gen const=1
global XvarC ‘$Xvars const”
foreach x of varlist $XvarC      {
gen ‘x’wt=wij*‘x’
egen ‘x’mnwt=total(‘x’wt), by(bcode)
gen g_‘x’mnwt=gmai*‘x’mnwt
gen Xdif‘x’=‘x’-g_‘x’mnwt
gen wXdif_‘x’=nwt*Xdif‘x’
} global XZvarC “$XvarC wXdif_*”
matrix accum XZij=$XZvarC, noconstant
matrix list XZij
matrix Prod1=XZij[$nx+2..($nx+1)*2, 1..$nx+1]
matrix list Prod1
matrix invprod1=inv(Prod1)
matrix vecaccum YpZ=lnincpp wXdif_*, noconstant
matrix list YpZ
matrix prod2=YpZ’
matrix list prod2
matrix beta=invprod1*prod2
matrix list beta
/*————————————————————————*/
/*to generate variance of beta*/
matrix accum ZpZ=wXdif_*, noconstant
matrix list ZpZ
matrix Tinvprod1=invprod1’
matrix prodE=invprod1*ZpZ*Tinvprod1
matrix list prodE
preserve
sort bcode
collapse (sum) wXdif_*, by(bcode)
matrix accum C=wXdif_*, noconstant
matrix list C
restore
matrix prodN=invprod1*C*Tinvprod1
matrix list prodN
scalar vE=varE[1]
scalar vU=varU[1]
matrix sum1=vE*prodE
matrix list sum1
forvalues i=1/$nx      {
display sqrt(sum1[‘i’,‘i’])
}
matrix sum2=vU*prodN
matrix list sum2
forvalues i=1/$nx      {
display sqrt(sum2[‘i’,‘i’])
}
matrix covB=sum1+sum2
matrix list covB
global np=$nx+1
forvalues i=1/$np      {
display sqrt(covB[‘i’,‘i’])
}
//generation of variance components
```

```
//MUNICIPAL LEVEL
sort bcode egen bcode1=group(bcode)
gen mcode=int(bcode/1000)
sort mcode egen mcode1=group(mcode)
by bcode1, sort: gen numhb=_N
by mcode1, sort: gen numhmun=_N
preserve collapse bcode1, by(mcode1)
global Nmun=_N
restore
preserve
collapse (mean) famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong
wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed
all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg
Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52 const, by(mcode) mkmat $XvarC,
matrix(X_barmun)
restore
matrix xVxmun=X_barmun*covB*X_barmun’
forvalues i=1/$Nmun        {
mkmat $XvarC if mcode1==‘i’, matrix(Xmun‘i’)
mkmat const if mcode1==‘i’, matrix(consmun‘i’)
matrix xVxmun[‘i’,1]=xVxmun[‘i’,1]+consmun‘i”*Xmun‘i’*covB*Xmun‘i”*consmun‘i’
}
//convert the matrix to single obs and generate the xcovBx data
svmat xVxmun, name(xVxmun)
preserve
keep xVxmun1
keep if xVxmun1< .
gen mcode1=_n
save xVxmunYR, replace
restore
//to generate number of clusters per municipality
preserve
collapse mcode, by(bcode)
by mcode, sort: gen Snclusmun=_N
collapse Snclusmun, by(mcode) save nclusmunYR, replace
restore
//to generate file containing number of hh
preserve
collapse mcode1 nhh=numhmun, by(mcode)
save mcodenhYR, replace
restore
//PROVINCIAL LEVEL
sort prov
egen prov1=group(prov)
```

```
by prov1, sort: gen numhprov=_N
preserve
collapse bcode1, by(prov1)
global Nprov=_N
restore
preserve
collapse (mean) famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong
wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed
all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg
Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52 const, by(prov1)
mkmat $XvarC, matrix(X_barprov)
restore
matrix xVxprov=X_barprov*covB*X_barprov’
/* forvalues i=1/$Nprov        {
mkmat $XvarC if prov1==‘i’, matrix(Xprov‘i’)
mkmat const if prov1==‘i’, matrix(consprov‘i’)
matrix xVxprov[‘i’,1]=xVxprov[‘i’,1]+consprov‘i”*Xprov‘i’*covB*Xprov‘i”*consprov‘i’
}*/
//convert the matrix to single obs and generate the xVx data
svmat xVxprov, name(xVxprov)
preserve
keep xVxprov1
keep if xVxprov1< .
gen prov1=_n save xVxprovYR, replace
restore
//to generate number of clusters per province
preserve
collapse prov, by(bcode)
by prov, sort: gen Snclusprov=_N
collapse Snclusprov, by(prov)
save nclusprovYR, replace
restore
//to generate file containing number of hh
preserve
collapse prov1 nhh=numhprov, by(prov)
save provnhYR, replace
restore
//REGIONAL LEVEL
sort regn egen reg1=group(regn)
by reg1, sort: gen numhreg=_N
preserve
collapse bcode1, by(reg1)
global Nreg=_N
restore
```

```
preserve
collapse (mean) famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong
wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed
all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg
Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52 const, by(reg1)
mkmat $XvarC, matrix(X_barreg)
restore
matrix xcovBxreg=X_barreg*covB*X_barreg’
matrix xVxreg=J($Nreg,1,0)
forvalues i=1/$Nreg       {
matrix xVxreg[‘i’,1]=xVxreg[‘i’,1]+xcovBxreg[‘i’,‘i’]
}
/*forvalues i=1/$Nreg       {
mkmat $XvarC if reg1==‘i’, matrix(Xreg‘i’)
mkmat const if reg1==‘i’, matrix(consreg‘i’) matrix xVxreg[‘i’,1]=xVxreg[‘i’,1]+consreg‘i”*Xreg‘i’*c
}*/
//convert the matrix to single obs and generate the xVx data
svmat xVxreg, name(xVxreg)
preserve
keep xVxreg1
keep if xVxreg1< .
gen reg1=_n
save xVxregYR, replace
restore
//to generate number of clusters per region
preserve
collapse regn reg1, by(bcode)
by reg1, sort: gen Snclusreg=_N
collapse Snclusreg, by(reg1)
save nclusregYR, replace
restore
//to generate file containing number of hh
preserve
collapse nhh=numhreg, by(reg1)
save regnhYR, replace
restore
//MUNICIPAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusmunYR”, clear
sort mcode
save nclusmunYR, replace
use “mcodenhYR”
sort mcode
merge mcode using “nclusmunYR”
```

```
drop _merge
save mcodenhYR, replace
//combine cluster, num HH and beta effect in one file
use “xVxmunYR”, replace
sort mcode1
save xVxmunYR, replace
use “mcodenhYR”
sort mcode1
merge mcode1 using “xVxmunYR”
gen beta_efctm=xVxmun1
drop _merge
save xVxmunYR, replace
//PROVINCIAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusprovYR”, clear
sort prov
save nclusprovYR, replace
use “provnhYR”
sort prov
merge prov using “nclusprovYR”
drop _merge
save provnhYR, replace
//combine cluster, num HH and beta effect in one file
use “xVxprovYR”, replace
sort prov1
save xVxprovYR, replace
use “provnhYR”
sort prov1
merge prov1 using “xVxprovYR”
gen beta_efctp=xVxprov1
drop _merge
save xVxprovYR, replace
//REGIONAL LEVEL COMBINING FILES
//combine cluster, num HH and beta effect in one file
use “nclusregYR”, clear
sort reg1 save nclusregYR, replace
use “regnhYR”
sort reg1
merge reg1 using “nclusregYR”
drop _merge
save regnhYR, replace
//combine cluster, num HH and beta effect in one file
use “xVxregYR”, replace
sort reg1
```

```
save xVxregYR, replace
use “regnhYR”
sort reg1
merge reg1 using “xVxregYR”
gen beta_efctp=xVxreg1
drop _merge
save xVxregYR, replace
/*to generate the Census clusterand hh effect as well as the file for variance component
estimates*/
use “E:\Size_bgy”, clear
gen nh2=nh*nh
gen nh2vU=(nh2*vU)
egen reg1=group(regn)
//MUNICIPAL LEVEL
preserve
by mcode, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(mcode)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctmunYR, replace
restore
//to generate varcomponents
preserve
use “xVxmunYR”, replace
sort mcode
save xVxmunYR, replace
use “CHHefctmunYR”
sort mcode
merge mcode using “xVxmunYR”
keep if _merge==3
drop _merge
save varcompsmunYR, replace
restore
//PROVINCIAL LEVEL
preserve
by prov, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(prov)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctprovYR, replace
restore
//to generate varcomponents
preserve
use “xVxprovYR”, replace
```

```
sort prov
save xVxprovYR, replace
use “CHHefctprovYR”
sort prov
merge prov using “xVxprovYR”
drop _merge
save varcompsprovYR, replace
restore
//REGIONAL LEVEL
preserve
by reg1, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(reg1)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh save CHHefctregYR, replace
restore
//to generate varcomponents
preserve
use “xVxregYR”, replace
sort reg1
save xVxregYR, replace
use “CHHefctregYR”
sort reg1
merge reg1 using “xVxregYR”
drop _merge
save varcompsregYR, replace
restore
```

IWEE.do

```
//This program is the implementation of the IWEE method.
clear
set mem 300m
set matsize 5000
cd “C:\SAE1\Results_YRK”
use “E:\SAE\SURVEY2_all.dta”, replace
/*Initial estimate of the components of variance*/
#delimit ;
global Xvars “famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong
wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed
all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg
Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52”;
#delimit cr
//to generate initial varEpsilon
global nx=0
```

```
foreach x of varlist $Xvars {
egen ‘x’mn=mean(‘x’), by(bcode)
gen new_‘x’=‘x’-‘x’mn
global nx=$nx+1
}
egen incmean=mean(lnincpp), by(bcode)
gen newinc=lnincpp-incmean
quietly regress newinc new_*
predict e_ij, resid
egen bn=group(bcode)
gen e2=e_ij*e_ij
egen SS1=total(e2)
//to generate the denominator for initial varEpsilon
by bcode, sort: gen bigN=_N
gen const=1
global XvarC “$Xvars const”
preserve
collapse (mean) bigN $XvarC, by(bcode)
gen id=_n
global num=_N
foreach x of varlist $XvarC {
gen new_‘x’=bigN*‘x’
}
matrix accum B = new_*, noconstant
display $num
restore
gen varE=(1/(_N-$num-($nx+1)+1))*SS1
//to generate initial VarUpsilon
matrix accum XpX = $Xvars
matrix invXX = syminv(XpX)
matrix F=invXX*B
matrix T=trace(F)
quietly
regress lnincpp $Xvars predict u_ij, resid
gen uij2=u_ij*u_ij
egen ssuij=total(uij2)
gen varU=(ssuij-((_N-($nx+1))*varE))/(_N-T[1,1])
display varU
display varE
//Some variables needed in the computation of beta and the variance components
egen wijtot=total(sswgthh), by(bcode)
gen wij=sswgthh/wijtot
gen wtlnincpp=wij*lnincpp
egen mnlnincp=total(wtlnincpp), by(bcode)
```

```
gen incmndif=lnincpp-mnlnincp
gen wij2=wij*wij
egen dltai2=total(wij2), by(bcode)
sort bcode
gen denomEps=(1-dltai2)*wijtot //denominator of weighted VarE
preserve
collapse denomEps, by(bcode)
egen TdenomEps=total(denomEps)
scalar TdenomEps=TdenomEps[1]
restore
foreach x of varlist $XvarC       {
gen ‘x’wt=wij*‘x’
egen ‘x’_mnwt=total(‘x’wt), by(bcode)
}
global Xvarmn ”*_mnwt”
foreach x of varlist $XvarC       {
gen dif_‘x’=‘x’-‘x’_mnwt
}
global Xvardif ”dif_*”
gen numerEps=1
gen viw=1 //initial values for step 4
gen viw2=1
gen qoutnt=1
gen gmai=varU/(varU+(varE*dltai2))
foreach x of varlist $XvarC       {
gen Z_‘x’=sswgthh*(‘x’-(gmai*‘x’_mnwt))
}
local XZvarC “$XvarC Z_*”
matrix accum XZ=‘XZvarC’, noconstant
matrix Prod1=XZ[$nx+2..($nx+1)*2, 1..$nx+1]
matrix vecaccum YpZ=lnincpp Z_*, noconstant
matrix beta=inv(Prod1)*YpZ’
gen varE1=0
gen varU1=0
//Estimation or generation of beta (Step1 of IWEE)
scalar crit1=1
scalar crit2=1
scalar crit3=1
while crit1>0.0001       { while crit2>0.0001       {       while crit3>0.0001       {
replace gmai=varU/(varU+(varE*dltai2))
foreach x of varlist $XvarC       {
replace Z_‘x’=sswgthh*(‘x’-(gmai*‘x’_mnwt))
}
local XZvarC “$XvarC Z_*”
```

```
matrix accum XZ=‘XZvarC’, noconstant
matrix Prod1=XZ[$nx+2..($nx+1)*2, 1..$nx+1]
matrix vecaccum YpZ=lnincpp Z_*, noconstant
matrix beta1=inv(Prod1)*YpZ’
matrix difbeta=beta1-beta
matrix Tdifbeta=difbeta’*difbeta
scalar crit1=Tdifbeta[1,1]
matrix beta=beta1
//Calculation of weighted varE to replace the Henderson estimate (Step 2 IWEE)
gen xb=0
scalar k=1
foreach x of varlist $Xvardif      {
replace xb=xb+beta[k,1]*‘x’
scalar k=k+1
}
replace numerEps=sswgthh*(incmndif-xb)*(incmndif-xb)
egen TnumerEps=total(numerEps)
replace varE1=TnumerEps/TdenomEps
scalar vE1=varE1[1]
scalar vE=varE[1]
scalar crit2=vE1-vE
replace varE=varE1
display varE
//Calculation of viw (Step 3 of IWEE)
scalar kl=1
gen xbar_b=0
foreach x of varlist $Xvarmn      {
replace xbar_b=xbar_b+beta[kl,1]*‘x’
scalar kl=kl+1
}
replace viw=gmai*(mnlnincp-xbar_b)
//Calculation of VarUpsilon (Step 4 of IWEE)
replace viw2=viw*viw preserve
collapse viw2, by(bcode)
egen Mnviw2=mean(viw2)
scalar Mnviw2=Mnviw2[1]
restore
replace qoutnt=(varE*varU*dltai2)/(varU+varE*dltai2)
preserve
collapse qoutnt, by(bcode)
egen Mnqoutnt=mean(qoutnt)
scalar Mnqoutnt=Mnqoutnt[1]
restore
replace varU1=Mnviw2+Mnqoutnt
```

```
scalar vU1=varU1[1]
scalar vU=varU[1]
scalar crit3=vU1-vU
replace varU=varU1
display varU
drop xbar_b
drop xb drop TnumerEps
}
}
}
//to generate VARIANCE OF BETA
scalar vE=varE[1]
scalar vU=varU[1]
matrix accum ZpZ=Z_*, noconstant
matrix prodE=inv(Prod1)*ZpZ*(inv(Prod1))’
preserve
sort bcode collapse (sum) Z_*, by(bcode)
matrix accum C=Z_*, noconstant
restore
matrix prodN=inv(Prod1)*C*(inv(Prod1))’
matrix covB=(vE*prodE)+(vU*prodN)
matrix list covB
scalar list
matrix list beta
global np=$nx+1
forvalues i=1/$np        {
display sqrt(covB[‘i’,‘i’])
}
svmat beta, name(beta)
preserve
keep beta1
keep if beta1< .
save beta_YRK, replace
restore
svmat covB, name(Var)
preserve
keep Var1-Var31
keep if Var1< .
save covBYRK, replace
restore
//just to clear some variables created
drop *wt *mn new_* Z_* *_mnwt dif_* wij wij wijtot wij2 u_ij gmai viw viw2 denomEps
numerEps
//generation of variance components
```

```
//MUNICIPAL LEVEL
sort bcode egen bcode1=group(bcode)
gen mcode=int(bcode/1000)
sort mcode egen mcode1=group(mcode)
by bcode1, sort: gen numhb=_N
by mcode1, sort: gen numhmun=_N
preserve
collapse bcode1, by(mcode1)
global Nmun=_N
restore
matrix xVxmun=J($Nmun,1,0)
forvalues i=1/$Nmun      {
mkmat $XvarC if mcode1==‘i’, matrix(Xmun‘i’)
mkmat const if mcode1==‘i’, matrix(consmun‘i’)
matrix xVxmun[‘i’,1]=xVxmun[‘i’,1]+consmun‘i”*Xmun‘i’*covB*Xmun‘i”*consmun‘i’
}
//convert the matrix to single obs and generate the xcovBx data
svmat xVxmun, name(xVxmun)
preserve
keep xVxmun1
keep if xVxmun1< .
gen mcode1=_n
save xVxmunYRK, replace
restore
//to generate number of clusters per municipality
preserve
collapse mcode, by(bcode)
by mcode, sort: gen Snclusmun=_N
collapse Snclusmun, by(mcode)
save nclusmunYRK, replace
restore
//to generate file containing number of hh
preserve
collapse mcode1 nhh=numhmun, by(mcode)
save mcodenhYRK, replace
restore
//PROVINCIAL LEVEL
sort prov
egen prov1=group(prov)
by prov1, sort: gen numhprov=_N
preserve
collapse bcode1, by(prov1)
global Nprov=_N
restore
```

```
matrix xVxprov=J($Nprov,1,0)
forvalues i=1/$Nprov        {
mkmat $XvarC if prov1==‘i’, matrix(Xprov‘i’)
mkmat const if prov1==‘i’, matrix(consprov‘i’)
matrix xVxprov[‘i’,1]=xVxprov[‘i’,1]+consprov‘i”*Xprov‘i’*covB*Xprov‘i”*consprov‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxprov, name(xVxprov)
preserve
keep xVxprov1
keep if xVxprov1< .
gen prov1=_n
save xVxprovYRK, replace
restore
//to generate number of clusters per province
preserve
collapse prov, by(bcode)
by prov, sort: gen Snclusprov=_N
collapse Snclusprov, by(prov)
save nclusprovYRK, replace
restore
//to generate file containing number of hh
preserve
collapse prov1 nhh=numhprov, by(prov)
save provnhYRK, replace
restore
//REGIONAL LEVEL
sort regn egen reg1=group(regn)
by reg1, sort: gen numhreg=_N
preserve
collapse bcode1, by(reg1)
global Nreg=_N restore
matrix xVxreg=J($Nreg,1,0)
forvalues i=1/$Nreg        {
mkmat $XvarC if reg1==‘i’, matrix(Xreg‘i’)
mkmat const if reg1==‘i’, matrix(consreg‘i’)
matrix xVxreg[‘i’,1]=xVxreg[‘i’,1]+consreg‘i”*Xreg‘i’*covB*Xreg‘i”*consreg‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxreg, name(xVxreg)
preserve
keep xVxreg1
keep if xVxreg1< .
gen reg1=_n
```

```
save xVxregYRK, replace
restore
//to generate number of clusters per region
preserve
collapse regn reg1, by(bcode)
by reg1, sort: gen Snclusreg=_N
collapse Snclusreg, by(reg1)
save nclusregYRK, replace
restore
//to generate file containing number of hh
preserve
collapse nhh=numhreg, by(reg1)
save regnhYRK, replace
restore
//MUNICIPAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusmunYRK”, clear
sort mcode
save nclusmunYRK, replace
use “mcodenhYRK”
sort mcode
merge mcode using “nclusmunYRK”
drop _merge
save mcodenhYRK, replace
//combine cluster, num HH and beta effect in one file
use “xVxmunYRK”, replace
sort mcode1
save xVxmunYRK, replace
use “mcodenhYRK”
sort mcode1
merge mcode1 using “xVxmunYRK”
gen beta_efctm=xVxmun1/(nhh*nhh)
drop _merge
save xVxmunYRK, replace
//PROVINCIAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusprovYRK”, clear
sort prov
save nclusprovYRK, replace
use “provnhYRK”
sort prov
merge prov using “nclusprovYRK”
drop _merge
save provnhYRK, replace
```

```
//combine cluster, num HH and beta effect in one file
use “xVxprovYRK”, replace
sort prov1
save xVxprovYRK, replace
use “provnhYRK”
sort prov1
merge prov1 using “xVxprovYRK”
gen beta_efctp=xVxprov1/(nhh*nhh)
drop _merge
save xVxprovYRK, replace
//REGIONAL LEVEL COMBINING FILES
//combine cluster, num HH and beta effect in one file
use “nclusregYRK”, clear
sort reg1 save nclusregYRK, replace
use “regnhYRK”
sort reg1
merge reg1 using “nclusregYRK”
drop _merge
save regnhYRK, replace
//combine cluster, num HH and beta effect in one file
use “xVxregYRK”, replace
sort reg1
save xVxregYRK, replace
use “regnhYRK”
sort reg1
merge reg1 using “xVxregYRK”
gen beta_efctp=xVxreg1/(nhh*nhh)
drop _merge
save xVxregYRK, replace
/*to generate the Census clusterand hh effect as well as the file for variance component
estimates*/
use “E:\Size_bgy”, clear
gen nh2=nh*nh gen nh2vU=(nh2*vU)
egen reg1=group(regn)
//MUNICIPAL LEVEL (mcode should be used as some mcodes in the survey do not
match with the census)
preserve
by mcode, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(mcode)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctmunYRK, replace
restore
//to generate varcomponents
```

```
preserve
use “xVxmunYRK”, replace
sort mcode
save xVxmunYRK, replace
use “CHHefctmunYRK”
sort mcode
merge mcode using “xVxmunYRK”
keep if _merge==3
drop _merge
save varcompsmunYRK, replace
restore
//PROVINCIAL LEVEL (prov or prov1 can be used as it poses no problem)
preserve
by prov, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(prov)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctprovYRK, replace
restore
//to generate varcomponents
preserve
use “xVxprovYRK”, replace
sort prov
save xVxprovYRK, replace
use “CHHefctprovYRK”
sort prov
merge prov using “xVxprovYRK”
drop _merge
save varcompsprovYRK, replace
restore
//REGIONAL LEVEL
preserve
by reg1, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(reg1)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh save CHHefctregYRK, replace
restore
//to generate varcomponents
preserve use “xVxregYRK”, replace
sort reg1 save xVxregYRK, replace
use “CHHefctregYRK”
sort reg1
merge reg1 using “xVxregYRK”
drop _merge
```

```
save varcompsregYRK, replace
restore
```

GSR.do

```
//This program is the implementation of the general survey regression method
clear
set mem 300m
set matsize 5000
cd “C:\SAE1”
use “E:\SAE\SURVEY2_all.dta”, replace
#delimit ;
global Xvars “famsize famsizesqc type_mult per_kids roof_light per_61up roof_strong
wall_light wall_salvaged wall_strong fa_xs fa_s fa_l fa_xl fa_xxl fa_xxxl all_eled all_hsed
all_coed dom_help head_male no_spouse hou_9600 hea_rel_mus Per_eng Hou_coelpg
Hou_own_ref Hou_own_tel Per_wor_prh Per_ind_52”;
#delimit cr
svyset bcode [pweight=sswgthh], strata(strata)
svy: regress lnincpp $Xvars
matrix covB=e(V)
preserve
svmat covB, name(VarB)
keep VarB1-VarB31
keep if VarB1< .
save covarB, replace
restore
gen const=1
global XvarC “$Xvars const”
//estimation of the variance of the small area estimate (municipal level)
//MUNICIPAL LEVEL
sort bcode egen bcode1=group(bcode)
gen mcode=int(bcode/1000)
sort mcode
egen mcode1=group(mcode)
by bcode1, sort: gen numhb=_N
by mcode1, sort: gen numhmun=_N
preserve
collapse bcode1, by(mcode1)
global Nmun=_N
restore
matrix xVxmun=J($Nmun,1,0)
forvalues i=1/$Nmun        {
mkmat $XvarC if mcode1==‘i’, matrix(Xmun‘i’)
mkmat const if mcode1==‘i’, matrix(consmun‘i’)
matrix xVxmun[‘i’,1]=xVxmun[‘i’,1]+consmun‘i”*Xmun‘i’*covB*Xmun‘i”*consmun‘i’
}
//convert the matrix to single obs and generate the xcovBx data
svmat xVxmun, name(xVxmun)
preserve
keep xVxmun1
keep if xVxmun1< .
gen mcode1=_n
save xVxmunGSR, replace
restore
//to generate number of clusters per municipality
preserve
collapse mcode, by(bcode)
by mcode, sort: gen Snclusmun=_N
collapse Snclusmun, by(mcode)
save nclusmunGSR, replace
restore
//to generate file containing number of hh
preserve
collapse mcode1 nhh=numhmun, by(mcode)
save mcodenhGSR, replace
restore
//PROVINCIAL LEVEL
sort prov
egen prov1=group(prov)
by prov1, sort: gen numhprov=_N
preserve
collapse bcode1, by(prov1)
global Nprov=_N
restore
matrix xVxprov=J($Nprov,1,0)
forvalues i=1/$Nprov      {
mkmat $XvarC if prov1==‘i’, matrix(Xprov‘i’)
mkmat const if prov1==‘i’, matrix(consprov‘i’)
matrix xVxprov[‘i’,1]=xVxprov[‘i’,1]+consprov‘i”*Xprov‘i’*covB*Xprov‘i”*consprov‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxprov, name(xVxprov)
preserve
keep xVxprov1
keep if xVxprov1< .
gen prov1=_n
save xVxprovGSR, replace
restore
```

```
//to generate number of clusters per province
preserve
collapse prov, by(bcode)
by prov, sort: gen Snclusprov=_N
collapse Snclusprov, by(prov)
save nclusprovGSR, replace
restore
//to generate file containing number of hh
preserve
collapse prov1 nhh=numhprov, by(prov)
save provnhGSR, replace
restore
//REGIONAL LEVEL
sort regn
egen reg1=group(regn)
by reg1, sort: gen numhreg=_N
preserve
collapse bcode1, by(reg1)
global Nreg=_N restore
matrix xVxreg=J($Nreg,1,0)
forvalues i=1/$Nreg        {
mkmat $XvarC if reg1==‘i’, matrix(Xreg‘i’)
mkmat const if reg1==‘i’, matrix(consreg‘i’)
matrix xVxreg[‘i’,1]=xVxreg[‘i’,1]+consreg‘i”*Xreg‘i’*covB*Xreg‘i”*consreg‘i’
}
//convert the matrix to single obs and generate the xVx data
svmat xVxreg, name(xVxreg)
preserve
keep xVxreg1
keep if xVxreg1< .
gen reg1=_n
save xVxregGSR, replace
restore
//to generate number of clusters per region
preserve
collapse regn reg1, by(bcode)
by reg1, sort: gen Snclusreg=_N
collapse Snclusreg, by(reg1)
save nclusregGSR, replace
restore
//to generate file containing number of hh
preserve
collapse nhh=numhreg, by(reg1)
save regnhGSR, replace
```

```
restore
//MUNICIPAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusmunGSR”, clear
sort mcode
save nclusmunGSR, replace
use “mcodenhGSR”
sort mcode
merge mcode using “nclusmunGSR”
drop _merge
save mcodenhGSR, replace
//combine cluster, num HH and beta effect in one file
use “xVxmunGSR”, replace
sort mcode1
save xVxmunGSR, replace
use “mcodenhGSR”
sort mcode1
merge mcode1 using “xVxmunGSR”
gen beta_efctm=xVxmun1/(nhh*nhh)
drop _merge
save xVxmunGSR, replace
//PROVINCIAL LEVEL COMBINING FILES
//combine number of cluster and number of HH in one file
use “nclusprovGSR”, clear
sort prov
save nclusprovGSR, replace
use “provnhGSR”
sort prov
merge prov using “nclusprovGSR”
drop _merge
save provnhGSR, replace
//combine cluster, num HH and beta effect in one file
use “xVxprovGSR”, replace
sort prov1
save xVxprovGSR, replace
use “provnhGSR”
sort prov1
merge prov1 using “xVxprovGSR”
gen beta_efctp=xVxprov1/(nhh*nhh)
drop _merge
save xVxprovGSR, replace
//REGIONAL LEVEL COMBINING FILES
//combine cluster, num HH and beta effect in one file
use “nclusregGSR”, clear
```

```
sort reg1 save nclusregGSR, replace
use “regnhGSR”
sort reg1
merge reg1 using “nclusregGSR”
drop _merge
save regnhGSR, replace
//combine cluster, num HH and beta effect in one file
use “xVxregGSR”, replace
sort reg1 save xVxregGSR, replace
use “regnhGSR”
sort reg1
merge reg1 using “xVxregGSR”
gen beta_efctp=xVxreg1/(nhh*nhh)
drop _merge
save xVxregGSR, replace
/*to generate the Census clusterand hh effect as well as the file for variance component
estimates*/
use “E:\Size_bgy”, clear
gen nh2=nh*nh
//variance come from the ELL method (not hetero)
gen varU=0.04741227
scalar vU=varU[1]
gen varE=0.18461
scalar vE=varE[1]
gen nh2vU=(nh2*vU)
egen reg1=group(regn)
//MUNICIPAL LEVEL
preserve
by mcode, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(mcode)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh save CHHefctmunGSR, replace
restore
//to generate varcomponents
preserve
use “xVxmunGSR”, replace
sort mcode
save xVxmunGSR, replace
use “CHHefctmunGSR”
sort mcode
merge mcode using “xVxmunGSR”
keep if _merge==3
drop _merge
save varcompsmunGSR, replace
```

```
restore
//PROVINCIAL LEVEL
preserve
by prov, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(prov)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh
save CHHefctprovGSR, replace
restore
//to generate varcomponents
preserve
use “xVxprovGSR”, replace
sort prov
save xVxprovGSR, replace
use “CHHefctprovGSR”
sort prov
merge prov using “xVxprovGSR”
drop _merge
save varcompsprovGSR, replace
restore
//REGIONAL LEVEL
preserve
by reg1, sort: gen Cnclus=_N
collapse Cnclus (sum) totnh=nh totnh2var=nh2vU, by(reg1)
gen clusfct=totnh2var/(totnh*totnh)
gen hhefct=vE/totnh save CHHefctregGSR, replace
restore
//to generate varcomponents
preserve
use “xVxregGSR”, replace
sort reg1
save xVxregGSR, replace
use “CHHefctregGSR”
sort reg1
merge reg1 using “xVxregGSR”
drop _merge
save varcompsregGSR, replace
restore
```

Appendix B

Regional and Provincial Level Models Fitted using ELL, PEB, IWEE and GSR

Table B.1: Region 1 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1893)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1233	0.0076	-0.1293	0.0069	-0.1238	0.0075	-0.1238	0.0075	-0.1179	0.0100
famsizesqc	0.0110	0.0016	0.0119	0.0015	0.0110	0.0016	0.0110	0.0016	0.0103	0.0020
dom_help	0.8104	0.0887	0.7562	0.1099	0.8073	0.0878	0.8071	0.0875	0.8449	0.0891
wall_light	-0.0681	0.0429	-0.0639	0.0374	-0.0602	0.0427	-0.0597	0.0426	-0.1447	0.0423
wall_strong	0.1376	0.0374	0.1521	0.0347	0.1451	0.0374	0.1456	0.0373	0.0612	0.0425
fa_xs	-0.2207	0.0491	-0.2237	0.0452	-0.2272	0.0488	-0.2276	0.0486	-0.1486	0.0567
fa_s	-0.1354	0.0384	-0.1225	0.0334	-0.1377	0.0380	-0.1379	0.0379	-0.1106	0.0454
fa_l	0.0948	0.0371	0.0889	0.0343	0.0959	0.0368	0.0960	0.0366	0.0853	0.0412
fa_xl	0.1663	0.0432	0.1552	0.0407	0.1694	0.0428	0.1696	0.0427	0.1370	0.0490
fa_xxl	0.3371	0.0455	0.3120	0.0483	0.3417	0.0452	0.3420	0.0450	0.2916	0.0515
fa_xxxl	0.3310	0.0619	0.3038	0.0603	0.3376	0.0613	0.3380	0.0611	0.2605	0.0663
all_hsed	0.3399	0.0525	0.3559	0.0478	0.3381	0.0521	0.3380	0.0519	0.3578	0.0484
all_coed	1.2182	0.0573	1.2476	0.0584	1.2079	0.0569	1.2073	0.0567	1.3298	0.0623
per_kids	-0.2470	0.0644	-0.2405	0.0585	-0.2444	0.0637	-0.2442	0.0635	-0.2742	0.0705
per_61up	-0.1461	0.0613	-0.1594	0.0579	-0.1470	0.0606	-0.1471	0.0604	-0.1352	0.0712
hou_9600	1.1398	0.4910	1.2704	0.4789	1.1432	0.5214	1.1436	0.5217	1.0751	0.5194
Hou_own_ref	1.4523	0.2455	1.5102	0.2386	1.4499	0.2607	1.4498	0.2609	1.4478	0.2359
const	9.3688	0.2032	9.3236	0.1966	9.3660	0.2150	9.3657	0.2151	9.4138	0.2143
Variance Components Estimate										
Cluster level		0.0307		0.0307		0.0373		0.0375		
Household level		0.1954		0.1893*		0.1905		0.1890		

Table B.2: Region 2 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1692)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1276	0.0091	-0.1237	0.0080	-0.1277	0.0092	-0.1279	0.0090	-0.1233	0.0109
famsizesqc	0.0122	0.0021	0.0114	0.0018	0.0122	0.0021	0.0122	0.0020	0.0118	0.0027
dom_help	0.6998	0.1232	0.6525	0.2403	0.6984	0.1245	0.6967	0.1222	0.7665	0.2165
wall_light	-0.0416	0.0476	-0.0408	0.0394	-0.0414	0.0481	-0.0408	0.0473	-0.0588	0.0569
wall_strong	0.1326	0.0439	0.1318	0.0390	0.1325	0.0444	0.1326	0.0437	0.1288	0.0550
fa_xs	-0.2085	0.0434	-0.2149	0.0348	-0.2096	0.0438	-0.2106	0.0431	-0.1695	0.0457
fa_s	-0.0217	0.0402	-0.0304	0.0339	-0.0221	0.0406	-0.0231	0.0399	0.0154	0.0439
fa_l	0.2743	0.0441	0.2261	0.0401	0.2739	0.0446	0.2736	0.0437	0.2886	0.0589
fa_xl	0.3297	0.0545	0.2675	0.0506	0.3296	0.0550	0.3294	0.0540	0.3414	0.0681
fa_xxl	0.4048	0.0570	0.2788	0.0517	0.4049	0.0576	0.4051	0.0565	0.4033	0.0784
fa_xxxl	0.3119	0.0682	0.2039	0.0686	0.3126	0.0689	0.3123	0.0677	0.3279	0.0933
all_hsed	0.3509	0.0582	0.3613	0.0484	0.3499	0.0588	0.3492	0.0577	0.3721	0.0660
all_coed	1.1807	0.0630	1.2810	0.0662	1.1771	0.0637	1.1737	0.0626	1.3020	0.0849
per_kids	-0.2139	0.0762	-0.2707	0.0674	-0.2131	0.0770	-0.2118	0.0756	-0.2582	0.0873
per_61up	-0.1900	0.0656	-0.1642	0.0584	-0.1899	0.0663	-0.1902	0.0651	-0.1792	0.0662
hou_9600	-0.2103	0.5818	-0.1671	0.5537	-0.2276	0.6006	-0.2266	0.6068	-0.2681	0.5973
Hou_own_ref	0.7954	0.3074	0.8120	0.2930	0.7957	0.3171	0.7981	0.3202	0.7068	0.3314
const	9.9783	0.2469	9.9765	0.2333	9.9844	0.2545	9.9846	0.2567	9.9764	0.2725
Variance Components Estimate										
Cluster level		0.0389		0.0389		0.0423		0.0443		
Household level		0.1901		0.1692*		0.1937		0.1863		

Table B.3: Region 3 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1478)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1159	0.0054	-0.1166	0.0049	-0.1159	0.0056	-0.1159	0.0054	-0.1152	0.0061
famsizesqc	0.0098	0.0011	0.0099	0.0010	0.0098	0.0012	0.0098	0.0011	0.0098	0.0011
dom_help	0.4543	0.0612	0.4614	0.0680	0.4536	0.0630	0.4500	0.0610	0.5311	0.0736
wall_light	-0.0080	0.0338	-0.0097	0.0279	-0.0080	0.0349	-0.0057	0.0338	-0.0538	0.0360
wall_strong	0.2570	0.0251	0.2441	0.0218	0.2569	0.0258	0.2571	0.0251	0.2516	0.0286
fa_xs	-0.1562	0.0335	-0.1556	0.0279	-0.1568	0.0345	-0.1581	0.0334	-0.1331	0.0384
fa_s	-0.1009	0.0305	-0.1043	0.0270	-0.1011	0.0314	-0.1016	0.0304	-0.0910	0.0307
fa_l	0.0593	0.0281	0.0615	0.0250	0.0594	0.0289	0.0599	0.0280	0.0529	0.0301
fa_xl	0.1526	0.0327	0.1577	0.0297	0.1526	0.0337	0.1534	0.0327	0.1419	0.0354
fa_xxl	0.1823	0.0321	0.1735	0.0298	0.1824	0.0330	0.1845	0.0321	0.1504	0.0320
fa_xxxl	0.3034	0.0368	0.2791	0.0339	0.3035	0.0379	0.3061	0.0368	0.2596	0.0361
all_hsed	0.3585	0.0349	0.3853	0.0318	0.3583	0.0360	0.3584	0.0348	0.3563	0.0348
all_coed	1.0982	0.0393	1.1812	0.0413	1.0975	0.0405	1.0951	0.0392	1.1462	0.0435
per_kids	-0.2284	0.0453	-0.2323	0.0399	-0.2287	0.0467	-0.2286	0.0452	-0.2283	0.0416
per_61up	-0.0452	0.0452	0.0038	0.0469	-0.0450	0.0466	-0.0452	0.0450	-0.0417	0.0504
hou_9600	-0.0660	0.2809	-0.0897	0.2719	-0.0671	0.2911	-0.0668	0.2940	-0.0683	0.3517
Hou_own_ref	1.1863	0.1595	1.2155	0.1541	1.1824	0.1653	1.1841	0.1669	1.1505	0.1632
const	9.8418	0.1356	9.8304	0.1302	9.8447	0.1404	9.8436	0.1412	9.8616	0.1610
Variance Components Estimate										
Cluster level		0.0284		0.0284		0.0307		0.0327		
Household level		0.1618		0.1478*		0.1716		0.1599		

Table B.4: Region 4 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1528)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1226	0.0044	-0.1228	0.0041	-0.1225	0.0045	-0.1227	0.0043	-0.1201	0.0051
famsizesqc	0.0097	0.0010	0.0103	0.0010	0.0097	0.0011	0.0097	0.0010	0.0094	0.0012
dom_help	0.6478	0.0442	0.6431	0.0543	0.6515	0.0451	0.6447	0.0436	0.7871	0.0824
wall_light	-0.1182	0.0249	-0.1038	0.0211	-0.1192	0.0255	-0.1163	0.0247	-0.1719	0.0293
wall_strong	0.1914	0.0198	0.1726	0.0172	0.1910	0.0202	0.1930	0.0196	0.1519	0.0244
fa_xs	-0.1988	0.0282	-0.1950	0.0241	-0.1977	0.0288	-0.1983	0.0279	-0.1880	0.0388
fa_s	-0.1068	0.0232	-0.1073	0.0195	-0.1061	0.0237	-0.1073	0.0229	-0.0829	0.0233
fa_l	0.0522	0.0210	0.0572	0.0182	0.0518	0.0215	0.0534	0.0208	0.0244	0.0214
fa_xl	0.1409	0.0239	0.1342	0.0212	0.1400	0.0244	0.1424	0.0236	0.0955	0.0296
fa_xxl	0.2720	0.0259	0.2679	0.0245	0.2710	0.0264	0.2737	0.0256	0.2216	0.0337
fa_xxxl	0.3502	0.0324	0.3306	0.0299	0.3487	0.0330	0.3530	0.0320	0.2696	0.0497
all_hsed	0.4207	0.0294	0.4443	0.0261	0.4225	0.0301	0.4187	0.0291	0.4989	0.0339
all_coed	1.1243	0.0322	1.2331	0.0330	1.1280	0.0329	1.1187	0.0318	1.3110	0.0459
per_kids	-0.2107	0.0373	-0.2155	0.0345	-0.2115	0.0382	-0.2101	0.0369	-0.2368	0.0439
per_61up	-0.0906	0.0359	-0.0546	0.0379	-0.0903	0.0367	-0.0905	0.0354	-0.0862	0.0451
hou_9600	0.5215	0.2265	0.5512	0.2199	0.5224	0.2261	0.5241	0.2315	0.4921	0.2046
Hou_own_ref	1.2441	0.1155	1.2929	0.1123	1.2444	0.1153	1.2486	0.1180	1.1642	0.1066
const	9.6576	0.1150	9.6113	0.1116	9.6572	0.1149	9.6553	0.1174	9.6899	0.1033
Variance Components Estimate										
Cluster level		0.0467		0.0467		0.0459		0.0502		
Household level		0.1713		0.1528*		0.1793		0.1665		

Table B.5: Region 5 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1578)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1185	0.0072	-0.1200	0.0063	-0.1184	0.0074	-0.1185	0.0072	-0.1138	0.0076
famsizesqc	0.0079	0.0016	0.0076	0.0012	0.0079	0.0016	0.0079	0.0016	0.0068	0.0016
dom_help	0.6551	0.0723	0.6311	0.0827	0.6558	0.0740	0.6547	0.0717	0.7091	0.0835
wall_light	-0.0952	0.0330	-0.0964	0.0257	-0.0953	0.0337	-0.0949	0.0328	-0.1129	0.0399
wall_strong	0.1393	0.0310	0.1255	0.0270	0.1391	0.0317	0.1398	0.0308	0.1078	0.0379
fa_xs	-0.1772	0.0335	-0.1724	0.0275	-0.1771	0.0343	-0.1770	0.0333	-0.1816	0.0418
fa_s	-0.0810	0.0313	-0.0676	0.0275	-0.0808	0.0321	-0.0810	0.0311	-0.0727	0.0367
fa_l	0.1095	0.0410	0.0917	0.0387	0.1094	0.0420	0.1098	0.0407	0.0888	0.0492
fa_xl	0.2596	0.0519	0.2388	0.0508	0.2592	0.0531	0.2598	0.0515	0.2315	0.0774
fa_xxl	0.2320	0.0544	0.1882	0.0579	0.2317	0.0556	0.2321	0.0540	0.2124	0.0816
fa_xxxl	0.3455	0.0674	0.2213	0.0710	0.3450	0.0690	0.3459	0.0669	0.3027	0.1100
all_hsed	0.3041	0.0507	0.3266	0.0447	0.3051	0.0519	0.3042	0.0504	0.3442	0.0533
all_coed	1.4816	0.0574	1.5833	0.0641	1.4836	0.0587	1.4816	0.0570	1.5685	0.0683
per_kids	-0.1973	0.0587	-0.2147	0.0510	-0.1975	0.0601	-0.1972	0.0583	-0.2113	0.0529
per_61up	-0.1380	0.0558	-0.1225	0.0474	-0.1376	0.0571	-0.1380	0.0553	-0.1168	0.0593
hou_9600	-0.3042	0.3261	-0.4001	0.3137	-0.2988	0.3296	-0.2986	0.3268	-0.3040	0.3923
Hou_own_ref	1.0291	0.2900	0.9311	0.2799	1.0372	0.2932	1.0384	0.2905	0.9877	0.3139
const	9.9206	0.1824	9.9850	0.1747	9.9158	0.1846	9.9157	0.1826	9.9155	0.2157
Variance Components Estimate										
Cluster level		0.0369		0.0369		0.0373		0.0374		
Household level		0.1709		0.1578*		0.1793		0.1684		

Table B.6: Region 6 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1659)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1271	0.0057	-0.1279	0.0052	-0.1272	0.0058	-0.1274	0.0057	-0.1206	0.0067
famsizesqc	0.0098	0.0013	0.0102	0.0011	0.0098	0.0013	0.0098	0.0012	0.0088	0.0014
dom_help	0.6010	0.0459	0.6575	0.0541	0.6019	0.0468	0.6003	0.0455	0.6706	0.0613
wall_light	-0.1048	0.0252	-0.1089	0.0208	-0.1052	0.0257	-0.1052	0.0251	-0.1056	0.0309
wall_strong	0.2357	0.0263	0.2056	0.0247	0.2351	0.0269	0.2352	0.0262	0.2267	0.0305
fa_xs	-0.1573	0.0308	-0.1624	0.0251	-0.1581	0.0315	-0.1590	0.0307	-0.1298	0.0342
fa_s	-0.0869	0.0280	-0.0878	0.0229	-0.0869	0.0286	-0.0869	0.0278	-0.0947	0.0282
fa_l	0.1118	0.0296	0.0875	0.0256	0.1120	0.0302	0.1129	0.0294	0.0770	0.0337
fa_xl	0.2001	0.0373	0.1656	0.0343	0.2001	0.0381	0.2017	0.0371	0.1336	0.0506
fa_xxl	0.2968	0.0379	0.2469	0.0385	0.2969	0.0386	0.2979	0.0376	0.2595	0.0499
fa_xxxl	0.3941	0.0487	0.3157	0.0482	0.3945	0.0496	0.3958	0.0483	0.3407	0.0770
all_hsed	0.3464	0.0401	0.3719	0.0341	0.3462	0.0408	0.3451	0.0398	0.3907	0.0414
all_coed	1.2134	0.0447	1.3136	0.0490	1.2121	0.0455	1.2092	0.0444	1.3275	0.0668
per_kids	-0.1905	0.0486	-0.2001	0.0451	-0.1898	0.0496	-0.1884	0.0483	-0.2513	0.0551
per_61up	-0.1283	0.0439	-0.1231	0.0416	-0.1286	0.0447	-0.1285	0.0435	-0.1305	0.0511
hou_9600	0.2588	0.3322	0.1663	0.3166	0.2690	0.3416	0.2699	0.3415	0.2252	0.3473
Hou_own_ref	1.2953	0.2010	1.2001	0.1933	1.2983	0.2063	1.2997	0.2062	1.2382	0.2056
const	9.6674	0.1679	9.7322	0.1597	9.6639	0.1725	9.6637	0.1722	9.6782	0.1752
Variance Components Estimate										
Cluster level		0.0387		0.0387		0.0413		0.0421		
Household level		0.1746		0.1659*		0.1813		0.1714		

Table B.7: Region 7 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1899)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1137	0.0075	-0.1172	0.0069	-0.1138	0.0077	-0.1140	0.0075	-0.1069	0.0082
famsizesqc	0.0101	0.0017	0.0105	0.0017	0.0101	0.0018	0.0102	0.0017	0.0090	0.0018
dom_help	0.5192	0.0649	0.5173	0.0640	0.5186	0.0659	0.5187	0.0644	0.5139	0.0710
wall_light	-0.1600	0.0351	-0.1592	0.0294	-0.1594	0.0357	-0.1594	0.0350	-0.1589	0.0395
wall_strong	0.1624	0.0318	0.1377	0.0282	0.1629	0.0323	0.1632	0.0317	0.1521	0.0419
fa_xs	-0.1925	0.0378	-0.1893	0.0320	-0.1943	0.0384	-0.1979	0.0376	-0.1128	0.0435
fa_s	-0.0876	0.0369	-0.0827	0.0319	-0.0884	0.0374	-0.0900	0.0366	-0.0508	0.0392
fa_l	0.1150	0.0397	0.1085	0.0350	0.1148	0.0403	0.1149	0.0394	0.1171	0.0492
fa_xl	0.1942	0.0530	0.1992	0.0478	0.1938	0.0538	0.1943	0.0526	0.1903	0.0595
fa_xxl	0.2259	0.0535	0.1946	0.0527	0.2263	0.0543	0.2275	0.0531	0.2029	0.0732
fa_xxxl	0.3640	0.0674	0.3730	0.0673	0.3647	0.0685	0.3656	0.0669	0.3495	0.0688
all_hsed	0.4688	0.0497	0.4647	0.0419	0.4682	0.0505	0.4649	0.0493	0.5504	0.0579
all_coed	1.3083	0.0557	1.4015	0.0616	1.3057	0.0566	1.3005	0.0554	1.4276	0.0939
per_kids	-0.1423	0.0625	-0.1685	0.0572	-0.1415	0.0634	-0.1389	0.0619	-0.2132	0.0717
per_61up	-0.1318	0.0550	-0.1318	0.0513	-0.1314	0.0558	-0.1306	0.0545	-0.1566	0.0594
hou_9600	1.9763	0.3473	2.0334	0.3329	1.9869	0.3579	1.9889	0.3662	1.9432	0.3331
Hou_own_ref	2.1120	0.1941	2.2311	0.1874	2.1202	0.1998	2.1237	0.2043	2.0448	0.1729
const	8.7933	0.1482	8.7787	0.1417	8.7896	0.1523	8.7907	0.1552	8.7649	0.1329
Variance Components Estimate										
Cluster level		0.0499		0.0499		0.0538		0.0582		
Household level		0.2131		0.1899*		0.2195		0.2086		

Table B.8: Region 8 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1757)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1176	0.0074	-0.1223	0.0069	-0.1173	0.0079	-0.1179	0.0074	-0.1106	0.0073
famsizesqc	0.0087	0.0017	0.0102	0.0016	0.0087	0.0018	0.0088	0.0016	0.0072	0.0015
dom_help	0.8091	0.0753	0.8350	0.0914	0.8120	0.0800	0.8055	0.0748	0.8897	0.1028
wall_light	-0.0332	0.0361	-0.0360	0.0305	-0.0337	0.0382	-0.0328	0.0359	-0.0436	0.0415
wall_strong	0.2351	0.0351	0.2128	0.0318	0.2344	0.0372	0.2363	0.0349	0.2140	0.0452
fa_xs	-0.1290	0.0401	-0.1247	0.0331	-0.1287	0.0424	-0.1296	0.0399	-0.1154	0.0457
fa_s	-0.0640	0.0374	-0.0591	0.0310	-0.0646	0.0397	-0.0633	0.0372	-0.0824	0.0430
fa_l	0.0657	0.0393	0.0635	0.0343	0.0658	0.0417	0.0654	0.0390	0.0692	0.0516
fa_xl	0.1815	0.0472	0.1345	0.0425	0.1820	0.0501	0.1813	0.0469	0.1848	0.0600
fa_xxl	0.2675	0.0450	0.2395	0.0469	0.2679	0.0477	0.2673	0.0448	0.2702	0.0644
fa_xxxl	0.3935	0.0540	0.3049	0.0511	0.3942	0.0572	0.3931	0.0537	0.4031	0.0682
all_hsed	0.3649	0.0530	0.3866	0.0475	0.3661	0.0562	0.3634	0.0527	0.3972	0.0506
all_coed	1.4915	0.0554	1.6250	0.0626	1.4958	0.0587	1.4862	0.0551	1.6086	0.0644
per_kids	-0.1009	0.0594	-0.0852	0.0525	-0.1023	0.0631	-0.0994	0.0590	-0.1400	0.0671
per_61up	-0.1596	0.0524	-0.1051	0.0507	-0.1596	0.0556	-0.1598	0.0520	-0.1582	0.0550
hou_9600	-0.2482	0.3201	-0.2708	0.3058	-0.2544	0.3306	-0.2547	0.3319	-0.2514	0.2573
Hou_own_ref	1.1637	0.2452	1.1648	0.2391	1.1606	0.2532	1.1696	0.2541	1.0562	0.2760
const	9.6364	0.1481	9.6505	0.1415	9.6381	0.1534	9.6384	0.1530	9.6339	0.1400
Variance Components Estimate										
Cluster level		0.0346		0.0346		0.0359		0.0388		
Household level		0.1850		0.1757*		0.2090		0.1815		

Table B.9: Region 9 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1727)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1274	0.0091	-0.1303	0.0082	-0.1278	0.0088	-0.1280	0.0089	-0.1141	0.0115
famsizesqc	0.0099	0.0018	0.0091	0.0014	0.0100	0.0018	0.0100	0.0018	0.0084	0.0021
dom_help	0.5884	0.0864	0.6825	0.0847	0.5876	0.0842	0.5874	0.0850	0.6057	0.0909
wall_light	-0.0526	0.0451	-0.0409	0.0394	-0.0543	0.0442	-0.0551	0.0447	-0.0064	0.0561
wall_strong	0.2120	0.0435	0.1928	0.0396	0.2122	0.0426	0.2122	0.0431	0.2004	0.0595
fa_xs	-0.1957	0.0450	-0.1852	0.0386	-0.1943	0.0441	-0.1938	0.0446	-0.2324	0.0594
fa_s	-0.0380	0.0449	-0.0651	0.0369	-0.0390	0.0438	-0.0392	0.0443	-0.0193	0.0493
fa_l	0.0383	0.0526	0.0383	0.0453	0.0386	0.0513	0.0389	0.0518	0.0074	0.0711
fa_xl	0.0841	0.0723	0.0717	0.0664	0.0825	0.0705	0.0819	0.0712	0.1248	0.0805
fa_xxl	0.2508	0.0754	0.2483	0.0800	0.2501	0.0736	0.2496	0.0744	0.2779	0.0925
fa_xxxl	0.2956	0.0863	0.2227	0.0814	0.2935	0.0843	0.2925	0.0852	0.3595	0.1449
all_hsed	0.3824	0.0609	0.4141	0.0555	0.3784	0.0594	0.3767	0.0600	0.4924	0.0759
all_coed	1.2537	0.0710	1.3010	0.0746	1.2466	0.0695	1.2433	0.0702	1.4719	0.1008
per_kids	-0.2354	0.0757	-0.2199	0.0692	-0.2335	0.0738	-0.2328	0.0744	-0.2801	0.0893
per_61up	-0.0948	0.0725	-0.0819	0.0757	-0.0964	0.0706	-0.0971	0.0712	-0.0571	0.1023
hou_9600	0.1094	0.5642	0.0196	0.5477	0.0947	0.5872	0.0952	0.6097	0.0439	0.7150
Hou_own_ref	1.8966	0.4433	1.8360	0.4312	1.8600	0.4608	1.8627	0.4784	1.6680	0.4362
const	9.4768	0.3268	9.5332	0.3168	9.4891	0.3394	9.4899	0.3522	9.4423	0.3428
Variance Components Estimate										
Cluster level		0.0747		0.0747		0.0838		0.0913		
Household level		0.1937		0.1727*		0.1835		0.1866		

Table B.10: Region 10 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1815)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.0930	0.0082	-0.0965	0.0076	-0.0930	0.0083	-0.0930	0.0081	-0.0910	0.0104
famsizesqc	0.0027	0.0020	0.0024	0.0019	0.0027	0.0020	0.0027	0.0019	0.0035	0.0023
dom_help	0.5651	0.0677	0.5491	0.0620	0.5633	0.0685	0.5613	0.0670	0.6338	0.0736
wall_light	-0.0195	0.0393	-0.0035	0.0317	-0.0198	0.0397	-0.0200	0.0390	-0.0117	0.0433
wall_strong	0.2191	0.0355	0.1969	0.0298	0.2189	0.0360	0.2189	0.0353	0.2196	0.0453
fa_xs	-0.1140	0.0464	-0.1128	0.0375	-0.1141	0.0470	-0.1138	0.0461	-0.1259	0.0551
fa_s	-0.0656	0.0412	-0.0576	0.0342	-0.0662	0.0417	-0.0671	0.0408	-0.0365	0.0535
fa_l	0.1151	0.0437	0.1165	0.0374	0.1152	0.0442	0.1151	0.0433	0.1214	0.0528
fa_xl	0.1732	0.0542	0.1645	0.0462	0.1744	0.0548	0.1761	0.0537	0.1169	0.0669
fa_xxl	0.2398	0.0545	0.2555	0.0446	0.2407	0.0552	0.2424	0.0541	0.1867	0.0693
fa_xxxl	0.4312	0.0560	0.3894	0.0491	0.4322	0.0567	0.4341	0.0556	0.3722	0.0843
all_hsed	0.4276	0.0562	0.4497	0.0476	0.4260	0.0568	0.4242	0.0556	0.4891	0.0641
all_coed	1.3549	0.0596	1.4857	0.0602	1.3512	0.0603	1.3470	0.0591	1.5108	0.0798
per_kids	-0.2898	0.0689	-0.3107	0.0625	-0.2904	0.0696	-0.2906	0.0682	-0.2723	0.0752
per_61up	-0.0363	0.0686	-0.0346	0.0665	-0.0373	0.0694	-0.0383	0.0680	-0.0005	0.0892
hou_9600	0.9900	0.4951	0.9768	0.4711	0.9777	0.5102	0.9751	0.5151	1.0729	0.3701
Hou_own_ref	1.7717	0.2955	1.7222	0.2827	1.7652	0.3044	1.7662	0.3073	1.7256	0.2530
const	8.8796	0.2442	8.9125	0.2324	8.8881	0.2515	8.8905	0.2536	8.8009	0.2078
Variance Components Estimate										
Cluster level		0.0514		0.0514		0.0555		0.0579		
Household level		0.2154		0.1815*		0.2197		0.2101		

Table B.11: Region 11 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1906)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.0981	0.0081	-0.1029	0.0076	-0.0980	0.0082	-0.0982	0.0080	-0.0952	0.0085
famsizesqc	0.0069	0.0019	0.0069	0.0014	0.0070	0.0019	0.0070	0.0018	0.0070	0.0020
dom_help	0.5467	0.0672	0.5766	0.0623	0.5471	0.0680	0.5412	0.0661	0.6429	0.0587
wall_light	-0.0905	0.0399	-0.0760	0.0347	-0.0913	0.0403	-0.0847	0.0392	-0.2010	0.0542
wall_strong	0.1926	0.0382	0.1825	0.0351	0.1920	0.0386	0.1962	0.0376	0.1206	0.0454
fa_xs	-0.1772	0.0384	-0.1832	0.0332	-0.1773	0.0388	-0.1785	0.0378	-0.1530	0.0513
fa_s	-0.1086	0.0351	-0.1208	0.0315	-0.1082	0.0355	-0.1101	0.0345	-0.0740	0.0420
fa_l	0.1625	0.0445	0.1384	0.0418	0.1625	0.0450	0.1638	0.0437	0.1346	0.0496
fa_xl	0.1284	0.0543	0.0898	0.0524	0.1287	0.0550	0.1278	0.0534	0.1466	0.0700
fa_xxl	0.2108	0.0541	0.1841	0.0557	0.2112	0.0547	0.2114	0.0532	0.2045	0.0632
fa_xxxl	0.2828	0.0766	0.2830	0.0765	0.2827	0.0775	0.2826	0.0753	0.2854	0.0732
all_hsed	0.3045	0.0528	0.3387	0.0469	0.3058	0.0534	0.3013	0.0519	0.3869	0.0706
all_coed	1.2414	0.0594	1.2892	0.0626	1.2425	0.0601	1.2311	0.0584	1.4472	0.0841
per_kids	-0.4305	0.0697	-0.4379	0.0675	-0.4303	0.0705	-0.4303	0.0684	-0.4314	0.0776
per_61up	-0.2770	0.0728	-0.2793	0.0699	-0.2768	0.0737	-0.2779	0.0715	-0.2594	0.0764
hou_9600	0.4199	0.4015	0.2985	0.3876	0.4165	0.4032	0.4083	0.4163	0.5501	0.4142
Hou_own_ref	1.2390	0.2976	1.2175	0.2893	1.2254	0.2988	1.2275	0.3087	1.1843	0.3008
const	9.5781	0.2316	9.6539	0.2233	9.5822	0.2326	9.5849	0.2396	9.5335	0.2484
Variance Components Estimate										
Cluster level		0.0505		0.0505		0.0507		0.0567		
Household level		0.2068		0.1906*		0.2118		0.1985		

Table B.12: Region 12 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.2166)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1094	0.0091	-0.1117	0.0085	-0.1095	0.0095	-0.1097	0.0090	-0.1064	0.0110
famsizesqc	0.0065	0.0019	0.0080	0.0019	0.0065	0.0020	0.0065	0.0019	0.0068	0.0020
dom_help	0.8977	0.0887	0.9232	0.1064	0.8949	0.0923	0.8918	0.0879	0.9552	0.0924
wall_light	-0.0445	0.0419	-0.0550	0.0359	-0.0444	0.0436	-0.0439	0.0417	-0.0521	0.0467
wall_strong	0.1620	0.0378	0.1281	0.0361	0.1624	0.0394	0.1641	0.0377	0.1356	0.0402
fa_xs	-0.2245	0.0432	-0.2294	0.0354	-0.2253	0.0450	-0.2263	0.0431	-0.2061	0.0429
fa_s	-0.1475	0.0399	-0.1518	0.0339	-0.1474	0.0416	-0.1463	0.0397	-0.1692	0.0415
fa_l	0.1183	0.0437	0.1091	0.0405	0.1178	0.0455	0.1186	0.0434	0.0967	0.0457
fa_xl	0.1389	0.0618	0.0765	0.0636	0.1378	0.0644	0.1366	0.0613	0.1563	0.0815
fa_xxl	0.3689	0.0707	0.2735	0.0855	0.3681	0.0736	0.3669	0.0701	0.3848	0.1084
fa_xxxl	0.0983	0.0861	0.1118	0.0851	0.0991	0.0897	0.0976	0.0855	0.1218	0.1013
all_hsed	0.3225	0.0589	0.3284	0.0499	0.3213	0.0613	0.3173	0.0584	0.3958	0.0535
all_coed	1.2469	0.0632	1.2773	0.0705	1.2425	0.0657	1.2338	0.0627	1.4033	0.0797
per_kids	-0.2291	0.0771	-0.2137	0.0723	-0.2284	0.0802	-0.2275	0.0763	-0.2420	0.0846
per_61up	0.0731	0.0745	0.0898	0.0698	0.0730	0.0775	0.0730	0.0738	0.0710	0.0723
hou_9600	0.7735	0.3743	0.7489	0.3572	0.7799	0.3967	0.7798	0.3983	0.7824	0.3902
Hou_own_ref	1.2908	0.2631	1.2219	0.2549	1.2962	0.2770	1.3004	0.2780	1.2222	0.2463
const	9.2527	0.2024	9.2972	0.1935	9.2496	0.2140	9.2506	0.2141	9.2258	0.2109
Variance Components Estimate										
Cluster level		0.0401		0.0401		0.0460		0.0485		
Household level		0.2105		0.2166*		0.2274		0.2054		

Table B.13: Region 13 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1676)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1197	0.0050	-0.1184	0.0047	-0.1196	0.0049	-0.1196	0.0050	-0.1205	0.0066
famsizesqc	0.0092	0.0010	0.0104	0.0011	0.0092	0.0010	0.0092	0.0010	0.0095	0.0013
dom_help	0.5422	0.0336	0.5484	0.0373	0.5362	0.0328	0.5373	0.0333	0.6330	0.0522
wall_light	-0.1075	0.0718	-0.0696	0.0587	-0.1064	0.0703	-0.1065	0.0714	-0.1095	0.0768
wall_strong	0.1711	0.0222	0.1644	0.0188	0.1714	0.0218	0.1713	0.0221	0.1571	0.0243
fa_xs	-0.2473	0.0272	-0.2376	0.0232	-0.2478	0.0266	-0.2478	0.0270	-0.2412	0.0306
fa_s	-0.1102	0.0264	-0.1007	0.0227	-0.1101	0.0258	-0.1101	0.0262	-0.1057	0.0279
fa_l	0.0611	0.0277	0.0763	0.0241	0.0621	0.0271	0.0621	0.0275	0.0630	0.0284
fa_xl	0.1654	0.0331	0.1686	0.0296	0.1677	0.0324	0.1675	0.0329	0.1554	0.0377
fa_xxl	0.3464	0.0365	0.3524	0.0352	0.3463	0.0357	0.3465	0.0362	0.3635	0.0459
fa_xxxl	0.5255	0.0392	0.4956	0.0367	0.5228	0.0384	0.5232	0.0390	0.5422	0.0647
all_hsed	0.3438	0.0369	0.3631	0.0320	0.3428	0.0360	0.3428	0.0366	0.3516	0.0411
all_coed	0.9293	0.0366	0.9783	0.0332	0.9279	0.0357	0.9284	0.0363	0.9849	0.0435
per_kids	-0.1478	0.0418	-0.1568	0.0367	-0.1458	0.0408	-0.1461	0.0415	-0.1736	0.0426
per_61up	-0.0367	0.0510	-0.0393	0.0517	-0.0352	0.0499	-0.0354	0.0506	-0.0674	0.0704
hou_9600	-0.3995	0.3299	-0.3484	0.3138	-0.3884	0.3360	-0.3896	0.3385	-0.4436	0.3254
Hou_own_ref	1.7026	0.2585	1.6507	0.2449	1.7349	0.2639	1.7329	0.2658	1.5723	0.2454
const	9.8095	0.1677	9.8046	0.1581	9.7880	0.1710	9.7892	0.1723	9.8807	0.1494
Variance Components Estimate										
Cluster level		0.0304		0.0304		0.0337		0.0339		
Household level		0.2002		0.1676*		0.1906		0.1967		

Table B.14: Region 14 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1664)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1227	0.0080	-0.1252	0.0075	-0.1230	0.0080	-0.1232	0.0078	-0.1140	0.0105
famsizesqc	0.0084	0.0018	0.0088	0.0017	0.0084	0.0019	0.0084	0.0018	0.0084	0.0024
dom_help	0.5310	0.1035	0.5084	0.1080	0.5295	0.1039	0.5282	0.1011	0.5613	0.1245
wall_light	-0.0502	0.0546	-0.0456	0.0453	-0.0481	0.0550	-0.0460	0.0536	-0.1226	0.0758
wall_strong	0.1746	0.0441	0.1570	0.0385	0.1780	0.0445	0.1803	0.0435	0.0902	0.0627
fa_xs	-0.1773	0.0440	-0.1498	0.0388	-0.1810	0.0443	-0.1832	0.0432	-0.0929	0.0493
fa_s	-0.0512	0.0412	-0.0359	0.0352	-0.0513	0.0414	-0.0517	0.0402	-0.0460	0.0494
fa_l	0.0993	0.0466	0.1161	0.0402	0.1018	0.0468	0.1032	0.0455	0.0352	0.0521
fa_xl	0.2354	0.0516	0.2264	0.0447	0.2390	0.0518	0.2414	0.0505	0.1348	0.0657
fa_xxl	0.3897	0.0528	0.3751	0.0507	0.3938	0.0531	0.3965	0.0518	0.2918	0.0743
fa_xxxl	0.4634	0.0602	0.4920	0.0547	0.4690	0.0606	0.4719	0.0590	0.3648	0.0701
all_hsed	0.3839	0.0550	0.3859	0.0488	0.3819	0.0552	0.3807	0.0537	0.4431	0.0682
all_coed	1.1160	0.0569	1.2244	0.0581	1.1109	0.0571	1.1075	0.0556	1.2667	0.0903
per_kids	-0.2087	0.0666	-0.2695	0.0615	-0.2058	0.0668	-0.2039	0.0650	-0.2985	0.0813
per_61up	-0.1902	0.0592	-0.2016	0.0600	-0.1876	0.0594	-0.1865	0.0577	-0.2573	0.0804
hou_9600	0.1921	0.2895	0.1164	0.2803	0.1964	0.3067	0.1977	0.3092	0.1359	0.3181
Hou_own_ref	1.2493	0.1469	1.2228	0.1433	1.2402	0.1553	1.2415	0.1564	1.1655	0.1541
const	9.7828	0.1165	9.8284	0.1109	9.7805	0.1215	9.7786	0.1212	9.8597	0.1378
Variance Components Estimate										
Cluster level		0.0670		0.0670		0.0777		0.0804		
Household level		0.1890		0.1664*		0.1900		0.1795		

Table B.15: Region 15 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1071)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1309	0.0076	-0.1284	0.0059	-0.1309	0.0076	-0.1309	0.0074	-0.1217	0.0100
famsizesqc	0.0106	0.0017	0.0101	0.0013	0.0106	0.0017	0.0106	0.0017	0.0097	0.0020
dom_help	0.9005	0.1090	1.0001	0.2716	0.9001	0.1093	0.9003	0.1072	0.8756	0.3023
wall_light	-0.0784	0.0342	-0.0417	0.0266	-0.0772	0.0343	-0.0778	0.0337	-0.0388	0.0664
wall_strong	0.1607	0.0338	0.1291	0.0281	0.1631	0.0339	0.1634	0.0334	0.1255	0.0582
fa_xs	-0.1949	0.0384	-0.1435	0.0268	-0.1926	0.0385	-0.1924	0.0378	-0.2292	0.0476
fa_s	-0.1028	0.0270	-0.0717	0.0195	-0.1009	0.0271	-0.1006	0.0266	-0.1428	0.0369
fa_l	0.0382	0.0316	0.0209	0.0231	0.0384	0.0317	0.0386	0.0311	0.0255	0.0437
fa_xl	0.1428	0.0444	0.0948	0.0329	0.1439	0.0444	0.1446	0.0436	0.0579	0.0751
fa_xxl	0.2134	0.0477	0.1097	0.0456	0.2127	0.0478	0.2128	0.0469	0.1914	0.0800
fa_xxxl	0.1509	0.0596	0.0939	0.0430	0.1494	0.0596	0.1492	0.0585	0.1795	0.0797
all_hsed	0.0632	0.0459	0.1130	0.0336	0.0646	0.0460	0.0642	0.0452	0.0927	0.0822
all_coed	0.7843	0.0548	0.7452	0.0603	0.7865	0.0550	0.7847	0.0540	0.9802	0.0971
per_kids	-0.1143	0.0586	-0.1474	0.0464	-0.1130	0.0587	-0.1124	0.0576	-0.1837	0.0889
per_61up	-0.0029	0.0741	0.0301	0.0782	-0.0022	0.0742	-0.0024	0.0728	0.0239	0.1309
hou_9600	-0.1094	0.2266	-0.3759	0.2221	-0.1553	0.2283	-0.1548	0.2290	-0.1916	0.1966
Hou_own_ref	0.5172	0.3147	0.2936	0.3106	0.4931	0.3161	0.4943	0.3172	0.3659	0.2604
const	10.0748	0.1240	10.2009	0.1175	10.0942	0.1248	10.0942	0.1248	10.1147	0.1224
Variance Components Estimate										
Cluster level		0.0675		0.0675		0.0710		0.0720		
Household level		0.1157		0.1071*		0.1161		0.1116		

Table B.16: Region 16 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1633)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1158	0.0081	-0.1236	0.0073	-0.1158	0.0082	-0.1159	0.0080	-0.1142	0.0077
famsizesqc	0.0067	0.0015	0.0075	0.0015	0.0067	0.0015	0.0067	0.0015	0.0063	0.0016
dom_help	0.7272	0.0853	0.6808	0.0972	0.7258	0.0865	0.7250	0.0844	0.7657	0.0904
wall_light	-0.0819	0.0455	-0.0722	0.0384	-0.0818	0.0464	-0.0818	0.0455	-0.0776	0.0434
wall_strong	0.1318	0.0405	0.1098	0.0374	0.1330	0.0414	0.1335	0.0406	0.1119	0.0456
fa_xs	-0.1071	0.0420	-0.1243	0.0353	-0.1100	0.0428	-0.1113	0.0418	-0.0522	0.0503
fa_s	-0.0739	0.0393	-0.0796	0.0336	-0.0743	0.0399	-0.0745	0.0389	-0.0643	0.0393
fa_l	0.1214	0.0434	0.0980	0.0390	0.1212	0.0440	0.1211	0.0429	0.1271	0.0454
fa_xl	0.2678	0.0575	0.1748	0.0535	0.2685	0.0583	0.2687	0.0570	0.2555	0.0583
fa_xxl	0.3268	0.0529	0.2887	0.0512	0.3261	0.0537	0.3257	0.0525	0.3459	0.0651
fa_xxxl	0.2600	0.0624	0.2177	0.0589	0.2604	0.0635	0.2607	0.0621	0.2443	0.1049
all_hsed	0.4706	0.0541	0.4507	0.0477	0.4683	0.0549	0.4672	0.0536	0.5242	0.0522
all_coed	1.5083	0.0640	1.6358	0.0702	1.5031	0.0650	1.5003	0.0635	1.6434	0.0828
per_kids	-0.1162	0.0684	-0.0866	0.0616	-0.1159	0.0693	-0.1158	0.0677	-0.1190	0.0773
per_61up	0.0195	0.0658	-0.0092	0.0596	0.0203	0.0667	0.0207	0.0651	0.0007	0.0602
hou_9600	-0.1025	0.2642	-0.1070	0.2483	-0.1010	0.2782	-0.1007	0.2765	-0.1111	0.2431
Hou_own_ref	0.7916	0.2376	0.7819	0.2267	0.7947	0.2502	0.7961	0.2487	0.7158	0.2350
const	9.5520	0.1554	9.5986	0.1453	9.5522	0.1628	9.5526	0.1615	9.5303	0.1551
Variance Components Estimate										
Cluster level		0.0303		0.0303		0.0350		0.0352		
Household level		0.1759		0.1633*		0.1803		0.1716		

Table B.17: Province 1 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.2375)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1450	0.0175	-0.1489	0.0156	-0.1452	0.0179	-0.1449	0.0171	-0.1413	0.0097
famsizesqc	0.0090	0.0063	0.0124	0.0067	0.0091	0.0065	0.0090	0.0062	0.0085	0.0055
fa_xs	-0.4549	0.1126	-0.3816	0.1010	-0.4552	0.1149	-0.4546	0.1095	-0.4479	0.0718
fa_s	-0.2550	0.0976	-0.2653	0.0794	-0.2545	0.0995	-0.2555	0.0951	-0.2693	0.1198
wall_light	-0.2055	0.0945	-0.1474	0.0778	-0.2057	0.0965	-0.2058	0.0919	-0.2063	0.1070
all_hsed	0.4007	0.1643	0.3531	0.1448	0.4015	0.1673	0.4006	0.1601	0.3891	0.1585
all_coed	1.5411	0.1677	1.8202	0.1769	1.5429	0.1709	1.5429	0.1635	1.5439	0.2326
Hou_own_tel	3.4373	1.0270	3.2630	1.0582	3.4265	1.0622	3.4274	0.9871	3.4392	0.5733
Per_wor_prh	-1.1075	1.1933	-1.5801	1.2008	-1.1049	1.2327	-1.1056	1.1483	-1.1150	0.8729
const	10.0976	0.1480	10.0798	0.1279	10.0988	0.1517	10.0981	0.1435	10.0872	0.1373
Variance Components Estimate										
Cluster level		0.0187		0.0187		0.0208		0.0167		
Household level		0.2575		0.2375*		0.2668		0.2450		

Table B.18: Province 2 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1926)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1285	0.0111	-0.1289	0.0095	-0.1291	0.0113	-0.1291	0.0110	-0.1247	0.0118
famsizesqc	0.0075	0.0017	0.0075	0.0012	0.0074	0.0017	0.0074	0.0017	0.0076	0.0020
fa_xs	-0.4732	0.0731	-0.4095	0.0670	-0.4744	0.0747	-0.4744	0.0730	-0.4628	0.0583
fa_s	-0.3440	0.0522	-0.3115	0.0443	-0.3425	0.0533	-0.3425	0.0521	-0.3502	0.0501
wall_light	-0.1777	0.0664	-0.1919	0.0537	-0.1781	0.0680	-0.1781	0.0664	-0.1738	0.0509
all_hsed	0.5090	0.0881	0.5547	0.0749	0.5004	0.0895	0.5004	0.0874	0.5526	0.0965
all_coed	1.7427	0.1031	1.8270	0.1113	1.7200	0.1053	1.7198	0.1029	1.8572	0.1179
Hou_own_tel	3.8979	0.7058	3.8213	0.6873	3.9253	0.7749	3.9255	0.7572	3.7773	0.7240
Per_wor_prh	-5.1010	1.8935	-5.2548	1.8657	-5.1382	2.0804	-5.1386	2.0331	-4.8794	1.7541
const	9.8231	0.1214	9.8074	0.1136	9.8318	0.1295	9.8319	0.1265	9.7687	0.1233
Variance Components Estimate										
Cluster level		0.0161		0.0161		0.0229		0.0219		
Household level		0.2103		0.1926*		0.2145		0.2045		

Table B.19: Province 3 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1565)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1264	0.0130	-0.1401	0.0124	-0.1266	0.0130	-0.1266	0.0126	-0.1248	0.0153
famsizesqc	0.0062	0.0035	0.0088	0.0034	0.0064	0.0035	0.0064	0.0034	0.0043	0.0039
fa_xs	-0.1326	0.0686	-0.1514	0.0661	-0.1455	0.0691	-0.1453	0.0670	-0.0580	0.0857
fa_s	-0.1156	0.0737	-0.1240	0.0722	-0.1168	0.0738	-0.1168	0.0717	-0.1120	0.0825
wall_light	-0.1678	0.0638	-0.1470	0.0588	-0.1686	0.0639	-0.1686	0.0621	-0.1590	0.0651
all_hsed	0.5223	0.1126	0.4864	0.1006	0.5020	0.1124	0.5023	0.1091	0.6617	0.0857
all_coed	1.8277	0.1463	2.0524	0.1794	1.8040	0.1465	1.8044	0.1422	1.9952	0.2201
Hou_own_tel	1.6336	1.5546	1.6757	1.5335	1.6740	1.7387	1.6739	1.6849	1.6126	1.2570
Per_wor_prh	3.4486	2.7824	2.7656	2.6989	3.4183	3.0976	3.4173	3.0020	2.9437	1.9000
const	9.6526	0.1186	9.7262	0.1160	9.6653	0.1255	9.6651	0.1217	9.5899	0.1347
Variance Components Estimate										
Cluster level		0.0242		0.0242		0.0343		0.0322		
Household level		0.1556		0.1565*		0.1534		0.1447		

Table B.20: Province 4 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1970)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1213	0.0125	-0.1345	0.0105	-0.1216	0.0131	-0.1218	0.0123	-0.1155	0.0109
famsizesqc	0.0104	0.0034	0.0119	0.0022	0.0104	0.0035	0.0105	0.0033	0.0083	0.0028
fa_xs	-0.2474	0.1099	-0.2768	0.0945	-0.2506	0.1148	-0.2541	0.1085	-0.1741	0.1934
fa_s	-0.2475	0.0829	-0.1846	0.0532	-0.2482	0.0865	-0.2480	0.0814	-0.2565	0.0870
wall_light	-0.3168	0.0596	-0.2734	0.0476	-0.3165	0.0622	-0.3155	0.0585	-0.3418	0.0546
all_hsed	0.2151	0.1095	0.2390	0.0906	0.2152	0.1143	0.2146	0.1075	0.2261	0.0990
all_coed	1.1489	0.1199	1.4035	0.1320	1.1473	0.1252	1.1438	0.1180	1.2358	0.1601
Hou_own_tel	1.9890	0.5579	1.8292	0.5524	2.0075	0.5884	2.0096	0.5788	1.9471	0.5734
Per_wor_prh	-1.5070	1.1325	-1.0563	1.0664	-1.4015	1.1946	-1.4035	1.1760	-1.3408	0.6912
const	10.2109	0.1566	10.1666	0.1441	10.2004	0.1646	10.2018	0.1599	10.1675	0.0824
Variance Components Estimate										
Cluster level		0.0405		0.0405		0.0457		0.0457		
Household level		0.1819		0.1970*		0.1979		0.1749		

Table B.21: Province 5 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1960)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1008	0.0128	-0.1078	0.0107	-0.1008	0.0136	-0.1008	0.0125	-0.1012	0.0138
famsizesqc	0.0012	0.0038	0.0031	0.0031	0.0013	0.0040	0.0014	0.0037	0.0001	0.0036
fa_xs	-0.2292	0.0751	-0.2345	0.0576	-0.2284	0.0799	-0.2267	0.0738	-0.2549	0.0836
fa_s	-0.1921	0.0719	-0.1483	0.0626	-0.1931	0.0764	-0.1950	0.0705	-0.1701	0.0723
wall_light	-0.1789	0.0689	-0.1755	0.0533	-0.1771	0.0734	-0.1735	0.0679	-0.2174	0.0669
all_hsed	0.4894	0.1226	0.4480	0.0998	0.4884	0.1303	0.4860	0.1201	0.5161	0.0853
all_coed	1.6779	0.1209	1.7250	0.1352	1.6749	0.1286	1.6675	0.1187	1.7594	0.1416
Hou_own_tel	1.3831	0.6310	1.4264	0.6242	1.3954	0.6819	1.4116	0.6548	1.2018	0.6207
Per_wor_prh	-0.6825	1.4573	-0.4476	1.4108	-0.6807	1.5763	-0.6647	1.5162	-0.8836	1.8457
const	9.9587	0.1578	9.9625	0.1417	9.9578	0.1698	9.9558	0.1614	9.9879	0.1341
Variance Components Estimate										
Cluster level		0.0188		0.0188		0.0228		0.0225		
Household level		0.1862		0.1960*		0.2098		0.1774		

Table B.22: Province 6 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1837)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1375	0.0119	-0.1401	0.0100	-0.1372	0.0124	-0.1374	0.0117	-0.1335	0.0132
famsizesqc	0.0131	0.0032	0.0166	0.0031	0.0131	0.0033	0.0131	0.0031	0.0132	0.0026
fa_xs	-0.1309	0.0976	-0.0890	0.0736	-0.1255	0.1013	-0.1292	0.0956	-0.0642	0.0738
fa_s	-0.0349	0.0758	0.0165	0.0577	-0.0321	0.0788	-0.0341	0.0743	0.0001	0.0559
wall_light	-0.2665	0.0643	-0.2862	0.0534	-0.2712	0.0666	-0.2680	0.0629	-0.3241	0.0670
all_hsed	0.4703	0.1312	0.4519	0.0952	0.4744	0.1362	0.4715	0.1285	0.5223	0.1249
all_coed	1.7354	0.1289	1.8990	0.1328	1.7409	0.1336	1.7369	0.1262	1.8082	0.1464
Hou_own_tel	1.6002	0.9539	1.7490	0.9062	1.5706	0.9445	1.5895	0.9201	1.2584	0.6808
Per_wor_prh	1.2950	1.4249	1.3660	1.3302	1.3073	1.4053	1.3077	1.3727	1.3008	1.5066
const	9.8711	0.1661	9.8341	0.1541	9.8688	0.1662	9.8696	0.1608	9.8553	0.1526
Variance Components Estimate										
Cluster level		0.0268		0.0268		0.0241		0.0243		
Household level		0.1928		0.1837*		0.2095		0.1853		

Table B.23: Province 7 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1130)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1238	0.0111	-0.1361	0.0098	-0.1238	0.0105	-0.1237	0.0109	-0.1235	0.0137
famsizesqc	0.0076	0.0020	0.0094	0.0023	0.0076	0.0018	0.0076	0.0019	0.0076	0.0020
fa_xs	-0.1217	0.0620	-0.0964	0.0537	-0.1244	0.0596	-0.1199	0.0603	-0.1065	0.0823
fa_s	-0.1338	0.0533	-0.1086	0.0417	-0.1335	0.0507	-0.1338	0.0522	-0.1381	0.0578
wall_light	-0.1370	0.0500	-0.0991	0.0416	-0.1404	0.0476	-0.1348	0.0488	-0.1132	0.0472
all_hsed	0.2233	0.0992	0.3074	0.0867	0.2188	0.0941	0.2268	0.0972	0.2578	0.1000
all_coed	1.1819	0.1096	1.3230	0.1316	1.1737	0.1043	1.1887	0.1072	1.2477	0.1410
Hou_own_tel	1.9064	0.9060	2.0001	0.8325	1.8964	0.9079	1.9038	0.8615	1.9252	0.9382
Per_wor_prh	1.2324	0.4717	1.1123	0.4063	1.2281	0.4705	1.2272	0.4498	1.2307	0.3705
const	9.8356	0.0962	9.8451	0.0844	9.8393	0.0937	9.8323	0.0929	9.8070	0.1080
Variance Components Estimate										
Cluster level		0.0114		0.0114		0.0128		0.0095		
Household level		0.1277		0.1130*		0.1134		0.1237		

Table B.24: Province 8 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.2440)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1302	0.0161	-0.1370	0.0141	-0.1289	0.0163	-0.1295	0.0159	-0.1335	0.0152
famsizesqc	0.0169	0.0037	0.0189	0.0035	0.0168	0.0037	0.0168	0.0036	0.0173	0.0047
fa_xs	-0.2477	0.1014	-0.2066	0.0847	-0.2510	0.1026	-0.2495	0.1003	-0.2394	0.0909
fa_s	-0.1921	0.0909	-0.1500	0.0882	-0.1974	0.0919	-0.1951	0.0899	-0.1791	0.0962
wall_light	-0.1678	0.1034	-0.2044	0.0815	-0.1634	0.1042	-0.1653	0.1020	-0.1787	0.1127
all_hsed	0.3221	0.1323	0.3074	0.1122	0.3182	0.1336	0.3199	0.1306	0.3316	0.1213
all_coed	1.4478	0.1378	1.4874	0.1408	1.4463	0.1393	1.4470	0.1362	1.4511	0.1858
Hou_own_tel	-1.2074	0.8385	-0.8632	0.7824	-1.2070	0.8911	-1.2072	0.8514	-1.2074	0.7701
Per_wor_prh	0.0919	1.3322	0.1302	1.1538	0.0729	1.4148	0.0809	1.3522	0.1374	1.2232
const	10.7834	0.2249	10.7367	0.2048	10.7815	0.2371	10.7824	0.2273	10.7885	0.2450
Variance Components Estimate										
Cluster level		0.0069		0.0069		0.0109		0.0087		
Household level		0.2772		0.2440*		0.2795		0.2686		

Table B.25: Province 9 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1927)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1365	0.0194	-0.1318	0.0176	-0.1362	0.0193	-0.1369	0.0183	-0.1387	0.0242
famsizesqc	0.0016	0.0062	0.0020	0.0051	0.0018	0.0062	0.0014	0.0059	0.0005	0.0096
fa_xs	-0.5041	0.1599	-0.3945	0.1463	-0.5124	0.1598	-0.4964	0.1510	-0.4610	0.1772
fa_s	-0.2661	0.1205	-0.0995	0.1230	-0.2765	0.1205	-0.2570	0.1137	-0.2117	0.2170
wall_light	-0.5792	0.4320	-0.5210	0.1798	-0.5905	0.4302	-0.5673	0.4092	-0.5135	0.1551
all_hsed	0.2680	0.1921	0.4302	0.1756	0.2653	0.1914	0.2708	0.1819	0.2826	0.2043
all_coed	1.1926	0.2247	1.4918	0.2625	1.1916	0.2248	1.1887	0.2122	1.1823	0.3841
Hou_own_tel	22.1576	20.4188	24.4940	19.0100	22.3619	22.1148	21.8245	18.2318	20.6183	9.9858
Per_wor_prh	3.9948	2.1687	3.2660	1.9995	4.0210	2.3308	3.9962	1.9484	3.9411	2.2967
const	10.3873	0.2306	10.2588	0.2132	10.3850	0.2482	10.3886	0.2069	10.3971	0.3160
Variance Components Estimate										
Cluster level		0.0124		0.0124		0.0175		0.0083		
Household level		0.1629		0.1927*		0.1614		0.1462		

Table B.26: Province 10 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1775)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1159	0.0091	-0.1143	0.0093	-0.1157	0.0098	-0.1161	0.0090	-0.1117	0.0086
famsizesqc	0.0084	0.0025	0.0091	0.0027	0.0084	0.0027	0.0084	0.0025	0.0081	0.0025
fa_xs	-0.1633	0.0778	-0.1432	0.0755	-0.1589	0.0833	-0.1648	0.0775	-0.1055	0.0649
fa_s	-0.1577	0.0662	-0.1630	0.0676	-0.1595	0.0713	-0.1576	0.0656	-0.1820	0.0606
wall_light	-0.3205	0.0597	-0.3261	0.0521	-0.3216	0.0643	-0.3202	0.0593	-0.3350	0.0557
all_hsed	0.5325	0.0841	0.5694	0.0745	0.5364	0.0907	0.5299	0.0834	0.6066	0.0905
all_coed	1.3271	0.0781	1.4095	0.0841	1.3308	0.0841	1.3239	0.0776	1.4009	0.0933
Hou_own_tel	0.8764	0.5090	0.9273	0.4958	0.8501	0.5272	0.8601	0.5167	0.7589	0.4662
Per_wor_prh	0.4426	1.6475	0.1323	1.6170	0.4463	1.7042	0.4708	1.6738	0.2060	1.3661
const	10.1698	0.2358	10.1542	0.2323	10.1747	0.2444	10.1751	0.2393	10.1671	0.1819
Variance Components Estimate										
Cluster level		0.0276		0.0276		0.0282		0.0295		
Household level		0.1815		0.1775*		0.2117		0.1781		

Table B.27: Province 11 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1677)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1286	0.0091	-0.1297	0.0085	-0.1300	0.0089	-0.1303	0.0088	-0.1187	0.0107
famsizesqc	0.0095	0.0025	0.0089	0.0024	0.0096	0.0025	0.0096	0.0024	0.0093	0.0030
fa_xs	-0.3323	0.0539	-0.3151	0.0513	-0.3572	0.0540	-0.3630	0.0536	-0.1890	0.0760
fa_s	-0.1538	0.0512	-0.1369	0.0420	-0.1613	0.0507	-0.1632	0.0501	-0.1027	0.0530
wall_light	-0.0577	0.2091	-0.1033	0.1129	-0.0292	0.2060	-0.0228	0.2033	-0.2519	0.1327
all_hsed	0.4091	0.0851	0.4658	0.0738	0.3920	0.0838	0.3884	0.0827	0.5223	0.0894
all_coed	1.0789	0.0801	1.1965	0.0763	1.0481	0.0792	1.0412	0.0783	1.2843	0.1282
Hou_own_tel	0.4787	0.5246	0.5389	0.5019	0.4646	0.5912	0.4651	0.6047	0.4260	0.5463
Per_wor_prh	2.2864	1.6809	1.8201	1.5933	2.3974	1.8943	2.4076	1.9378	2.0734	1.7326
const	10.3303	0.1082	10.3197	0.1024	10.3482	0.1168	10.3533	0.1182	10.1787	0.0683
Variance Components Estimate										
Cluster level		0.0354		0.0354		0.0503		0.0538		
Household level		0.1842		0.1677*		0.1766		0.1715		

Table B.28: Province 12 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.2565)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1134	0.0121	-0.1188	0.0103	-0.1133	0.0124	-0.1133	0.0119	-0.1123	0.0122
famsizesqc	0.0193	0.0038	0.0159	0.0037	0.0192	0.0040	0.0192	0.0038	0.0182	0.0043
fa_xs	-0.3404	0.0787	-0.2383	0.0595	-0.3396	0.0810	-0.3397	0.0774	-0.3250	0.0716
fa_s	-0.1785	0.0692	-0.1021	0.0537	-0.1770	0.0711	-0.1772	0.0680	-0.1487	0.0691
wall_light	-0.1670	0.0695	-0.1571	0.0503	-0.1662	0.0714	-0.1663	0.0683	-0.1508	0.0467
all_hsed	0.5806	0.1206	0.6200	0.1035	0.5808	0.1243	0.5807	0.1188	0.5853	0.1343
all_coed	1.9139	0.1258	1.9218	0.1695	1.9154	0.1293	1.9152	0.1237	1.9482	0.1857
Hou_own_tel	2.5118	0.5291	2.5284	0.4895	2.5100	0.5360	2.5100	0.5134	2.5130	0.5037
Per_wor_prh	0.0936	0.6335	-0.1802	0.5649	0.0968	0.6413	0.0967	0.6143	0.1108	0.6262
const	9.4380	0.1070	9.4536	0.0923	9.4369	0.1092	9.4370	0.1046	9.4175	0.1300
Variance Components Estimate										
Cluster level		0.0158		0.0158		0.0154		0.0142		
Household level		0.2227		0.2565*		0.2369		0.2167		

Table B.29: Province 13 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1880)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1032	0.0139	-0.1159	0.0113	-0.1032	0.0145	-0.1032	0.0137	-0.1048	0.0228
famsizesqc	0.0041	0.0033	0.0062	0.0026	0.0041	0.0034	0.0040	0.0032	0.0051	0.0048
fa_xs	-0.2712	0.0856	-0.2444	0.0647	-0.2712	0.0891	-0.2696	0.0843	-0.3170	0.0775
fa_s	-0.2764	0.0838	-0.2312	0.0579	-0.2768	0.0872	-0.2765	0.0825	-0.2850	0.0944
wall_light	-0.2558	0.0742	-0.2202	0.0525	-0.2557	0.0772	-0.2567	0.0730	-0.2239	0.0867
all_hsed	0.4773	0.1251	0.4440	0.0952	0.4770	0.1303	0.4741	0.1232	0.5769	0.1182
all_coed	1.5930	0.1386	1.8085	0.1425	1.5930	0.1442	1.5885	0.1365	1.7554	0.1635
Hou_own_tel	0.4947	1.4170	0.4817	1.3452	0.5118	1.4647	0.5140	1.4204	0.4458	1.1535
Per_wor_prh	2.2662	2.0695	2.3766	1.9370	2.2538	2.1390	2.2624	2.0746	1.9554	1.7814
const	9.7674	0.1660	9.7827	0.1486	9.7691	0.1719	9.7695	0.1656	9.7534	0.1790
Variance Components Estimate										
Cluster level		0.0544		0.0544		0.0578		0.0555		
Household level		0.2622		0.1880*		0.2843		0.2535		

Table B.30: Province 14 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.1427)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1258	0.0080	-0.1306	0.0068	-0.1261	0.0085	-0.1263	0.0080	-0.1230	0.0089
famsizesqc	0.0093	0.0024	0.0098	0.0016	0.0094	0.0025	0.0095	0.0024	0.0082	0.0020
fa_xs	-0.3178	0.0537	-0.3006	0.0396	-0.3136	0.0567	-0.3109	0.0542	-0.3513	0.0595
fa_s	-0.2302	0.0453	-0.2268	0.0381	-0.2296	0.0476	-0.2292	0.0453	-0.2333	0.0423
wall_light	-0.1562	0.0791	-0.1450	0.0633	-0.1538	0.0835	-0.1520	0.0797	-0.1802	0.0536
all_hsed	0.2981	0.0629	0.3320	0.0545	0.3021	0.0662	0.3048	0.0630	0.2601	0.0641
all_coed	1.2410	0.0699	1.3471	0.0712	1.2381	0.0735	1.2366	0.0699	1.2605	0.0825
Hou_own_tel	0.2127	0.6714	0.0973	0.6391	0.2090	0.7717	0.2103	0.7837	0.1874	0.4777
Per_wor_prh	-0.2089	0.6938	-0.4503	0.6941	-0.1755	0.7984	-0.1758	0.8116	-0.1651	0.6101
const	10.7386	0.1715	10.7807	0.1627	10.7360	0.1957	10.7354	0.1978	10.7436	0.1255
Variance Components Estimate										
Cluster level		0.0245		0.0245		0.0358		0.0391		
Household level		0.1671		0.1427*		0.1831		0.1647		

Table B.31: Province 15 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1874)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1424	0.0121	-0.1444	0.0101	-0.1422	0.0129	-0.1424	0.0120	-0.1395	0.0142
famsizesqc	0.0127	0.0024	0.0145	0.0031	0.0127	0.0026	0.0127	0.0024	0.0128	0.0028
fa_xs	-0.3948	0.0699	-0.3811	0.0549	-0.3923	0.0738	-0.3949	0.0690	-0.3615	0.0858
fa_s	-0.1717	0.0609	-0.1249	0.0492	-0.1692	0.0643	-0.1715	0.0601	-0.1422	0.0812
wall_light	-0.1254	0.0610	-0.1123	0.0431	-0.1286	0.0645	-0.1254	0.0602	-0.1658	0.0598
all_hsed	0.4020	0.1059	0.4679	0.0806	0.4040	0.1122	0.4023	0.1046	0.4219	0.1256
all_coed	1.2850	0.1199	1.2716	0.1301	1.2878	0.1268	1.2845	0.1184	1.3239	0.1430
Hou_own_tel	0.7018	0.5600	0.9501	0.4973	0.6992	0.5629	0.7015	0.5527	0.6726	0.3252
Per_wor_prh	1.5150	1.1434	1.0893	1.0290	1.5290	1.1486	1.5313	1.1283	1.5044	1.0714
const	10.1599	0.1134	10.1359	0.1057	10.1583	0.1166	10.1601	0.1119	10.1365	0.1203
Variance Components Estimate										
Cluster level		0.0255		0.0255		0.0238		0.0248		
Household level		0.1851		0.1874*		0.2085		0.1805		

Table B.32: Province 16 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.2439)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1186	0.0163	-0.1188	0.0117	-0.1187	0.0173	-0.1187	0.0161	-0.1181	0.0146
famsizesqc	0.0122	0.0045	0.0116	0.0032	0.0125	0.0047	0.0127	0.0044	0.0112	0.0044
fa_xs	-0.4977	0.0956	-0.4196	0.0716	-0.4813	0.1029	-0.4732	0.0968	-0.5368	0.1019
fa_s	-0.2326	0.0796	-0.1704	0.0706	-0.2232	0.0851	-0.2186	0.0799	-0.2550	0.0731
wall_light	-0.0055	0.0919	-0.0496	0.0670	-0.0096	0.0978	-0.0116	0.0916	0.0021	0.0835
all_hsed	0.3839	0.1332	0.3773	0.1076	0.3785	0.1412	0.3758	0.1319	0.4032	0.1504
all_coed	2.0387	0.1634	2.0092	0.2014	2.0376	0.1741	2.0369	0.1631	2.0490	0.1681
Hou_own_tel	5.2064	1.4569	4.9280	1.4064	5.2296	1.7183	5.2410	1.7094	5.1403	0.8690
Per_wor_prh	-5.3469	2.5302	-4.9008	2.4615	-5.3569	2.9888	-5.3606	2.9756	-5.3341	1.3490
const	9.9553	0.1612	9.9204	0.1413	9.9504	0.1809	9.9478	0.1752	9.9655	0.1174
Variance Components Estimate										
Cluster level		0.0115		0.0115		0.0215		0.0240		
Household level		0.2144		0.2439*		0.2374		0.2058		

Table B.33: Province 17 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.1898)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1231	0.0117	-0.1234	0.0104	-0.1231	0.0118	-0.1231	0.0115	-0.1229	0.0133
famsizesqc	0.0071	0.0028	0.0064	0.0018	0.0071	0.0028	0.0071	0.0027	0.0070	0.0028
fa_xs	-0.3195	0.0636	-0.3017	0.0506	-0.3192	0.0643	-0.3194	0.0625	-0.2586	0.0765
fa_s	-0.1584	0.0586	-0.1344	0.0494	-0.1585	0.0593	-0.1586	0.0576	-0.1068	0.0670
wall_light	-0.2659	0.0607	-0.2271	0.0454	-0.2655	0.0614	-0.2656	0.0597	-0.2287	0.0732
all_hsed	0.3009	0.1119	0.3878	0.1030	0.3016	0.1132	0.3011	0.1100	0.4451	0.1213
all_coed	1.5833	0.1197	1.7482	0.1271	1.5851	0.1211	1.5845	0.1177	1.7662	0.1665
Hou_own_tel	1.7801	0.5712	1.5265	0.5379	1.7779	0.5802	1.7786	0.5657	1.5804	0.4906
Per_wor_prh	1.0845	1.7863	1.4126	1.6819	1.1204	1.8150	1.1194	1.7697	1.4522	1.1880
const	10.0002	0.1530	9.9438	0.1435	9.9963	0.1552	9.9967	0.1512	9.8904	0.1689
Variance Components Estimate										
Cluster level		0.0545		0.0545		0.0565		0.0539		
Household level		0.2095		0.1898*		0.2142		0.2021		

Table B.34: Province 18 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean=0.2034)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.0885	0.0169	-0.0925	0.0127	-0.0893	0.0168	-0.0888	0.0164	-0.0835	0.0210
famsizesqc	0.0131	0.0053	0.0108	0.0040	0.0131	0.0053	0.0131	0.0051	0.0126	0.0064
fa_xs	-0.2716	0.0881	-0.2296	0.0670	-0.2722	0.0877	-0.2721	0.0856	-0.2741	0.0680
fa_s	-0.3095	0.0871	-0.2852	0.0645	-0.3109	0.0866	-0.3098	0.0846	-0.2946	0.0864
wall_light	-0.2352	0.0805	-0.2774	0.0606	-0.2458	0.0808	-0.2387	0.0784	-0.1612	0.0650
all_hsed	0.3049	0.1604	0.3404	0.1303	0.2987	0.1594	0.3028	0.1557	0.3519	0.1667
all_coed	1.8039	0.1578	1.8068	0.1743	1.7920	0.1578	1.8000	0.1535	1.8823	0.2895
Hou_own_tel	0.5310	2.7621	-0.2275	2.5692	0.5096	3.0422	0.5081	2.7660	0.5132	3.1798
Per_wor_prh	1.7031	4.4569	-0.2550	4.2326	1.7688	4.9166	1.7774	4.4654	1.8612	3.9005
const	9.6622	0.1772	9.8002	0.1576	9.6733	0.1882	9.6652	0.1754	9.5760	0.1557
Variance Components Estimate										
Cluster level		0.0322		0.0322		0.0434		0.0335		
Household level		0.2095		0.2034*		0.2055		0.1970		

Table B.35: Province 19 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.2131)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1167	0.0138	-0.1059	0.0115	-0.1166	0.0139	-0.1166	0.0134	-0.1171	0.0151
famsizesqc	0.0096	0.0050	0.0086	0.0044	0.0097	0.0050	0.0097	0.0049	0.0076	0.0050
fa_xs	-0.2287	0.0776	-0.2449	0.0637	-0.2307	0.0789	-0.2306	0.0760	-0.2096	0.0685
fa_s	-0.1968	0.0733	-0.2105	0.0596	-0.1934	0.0741	-0.1934	0.0714	-0.2390	0.0645
wall_light	-0.3271	0.0620	-0.2249	0.0509	-0.3337	0.0628	-0.3337	0.0605	-0.2332	0.0715
all_hsed	0.5298	0.1395	0.5319	0.1209	0.5395	0.1413	0.5394	0.1361	0.3930	0.1433
all_coed	1.5023	0.1384	1.4858	0.1502	1.4969	0.1400	1.4970	0.1348	1.5964	0.1881
Hou_own_tel	4.4743	0.9467	3.4783	0.9084	4.4156	1.0331	4.4158	0.9937	4.5710	0.8346
Per_wor_prh	-3.0420	1.9205	-3.1354	1.7316	-2.9546	2.0970	-2.9548	2.0170	-3.2694	1.9220
const	9.9929	0.1518	9.9642	0.1366	9.9890	0.1612	9.9890	0.1551	9.9842	0.1685
Variance Components Estimate										
Cluster level		0.0444		0.0444		0.0563		0.0521		
Household level		0.2069		0.2131*		0.2104		0.1951		

Table B.36: Province 20 estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household(mean= 0.2797)

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1390	0.0186	-0.1286	0.0160	-0.1395	0.0188	-0.1394	0.0181	-0.1376	0.0276
famsizesqc	0.0137	0.0059	0.0145	0.0046	0.0138	0.0060	0.0138	0.0057	0.0133	0.0082
fa_xs	-0.4019	0.1087	-0.3512	0.0887	-0.3939	0.1109	-0.3956	0.1064	-0.4267	0.0832
fa_s	-0.2751	0.1169	-0.2353	0.0953	-0.2880	0.1193	-0.2853	0.1145	-0.2387	0.0820
wall_light	-0.1530	0.1218	-0.0971	0.1042	-0.1464	0.1241	-0.1478	0.1191	-0.1716	0.0756
all_hsed	0.6329	0.2158	0.4052	0.2169	0.6310	0.2198	0.6313	0.2110	0.6458	0.1812
all_coed	2.0501	0.1752	2.2096	0.1902	2.0722	0.1804	2.0675	0.1727	1.9903	0.2939
Hou_own_tel	1.1867	1.1201	0.9045	1.0860	1.1680	1.3100	1.1723	1.2174	1.2425	1.2183
Per_wor_prh	3.1624	2.7785	4.3866	2.7024	3.1187	3.2550	3.1277	3.0238	3.2791	1.7043
const	9.7209	0.1930	9.5911	0.1712	9.7225	0.2195	9.7222	0.2052	9.7153	0.1207
Variance Components Estimate										
Cluster level		0.0175		0.0175		0.0314		0.0257		
Household level		0.2345		0.2797*		0.2389		0.2209		

Appendix C

Variance-covariance Structure of SPREE Estimates

The following variance-covariance structures were derived by (Purcell, 1979).
Balanced Repeated Replicates:

$$V_{R1} = \frac{1}{2\tilde{H}} \sum_{\tilde{h}=1}^{\tilde{H}} \left[(\hat{\mathbf{p}}_{(1\tilde{h})} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(1\tilde{h})} - \hat{\mathbf{p}})' + (\hat{\mathbf{p}}_{(2\tilde{h})} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(2\tilde{h})} - \hat{\mathbf{p}})' \right] \tag{C.1}$$

$$V_{R2} = \frac{1}{\tilde{H}} \sum_{\tilde{h}=1}^{\tilde{H}} \left[(\hat{\mathbf{p}}_{(1\tilde{h})} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(1\tilde{h})} - \hat{\mathbf{p}})' \right] \tag{C.2}$$

$$V_{R3} = \frac{1}{4\tilde{H}} \sum_{\tilde{h}=1}^{\tilde{H}} \left[(\hat{\mathbf{p}}_{(1\tilde{h})} - \hat{\mathbf{p}}_{(2\tilde{h})})(\hat{\mathbf{p}}_{(1\tilde{h})} - \hat{\mathbf{p}}_{(2\tilde{h})})' \right] \tag{C.3}$$

where $\hat{\mathbf{p}}$ is the full sample estimates based on the survey margins derived from the full sample and $\hat{\mathbf{p}}_{(1\tilde{h})}$ is the set of ESPREE estimates based on the survey margins estimated from the $\tilde{h}$th half-sample, formed by including one of the two replicates from each stratum, while $\hat{\mathbf{p}}_{(2\tilde{h})}$ is the set of SPREE estimates based on the allocation structure estimated from the complement half-sample, formed by the replicates not in the $\tilde{h}$th half-sample. $\tilde{H}$ is the number of half-samples or replicates.

Jackknife:

$$V_{J1} = \frac{1}{2} \sum_{g=1}^{G} \left[(\hat{\mathbf{p}}_{(1g)} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(1g)} - \hat{\mathbf{p}})' + (\hat{\mathbf{p}}_{(2g)} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(2g)} - \hat{\mathbf{p}})' \right] \tag{C.4}$$

$$V_{J2} = \sum_{g=1}^{G} \left[(\hat{\mathbf{p}}_{(1g)} - \hat{\mathbf{p}})(\hat{\mathbf{p}}_{(1g)} - \hat{\mathbf{p}})' \right] \tag{C.5}$$

$$V_{J3} = \sum_{g=1}^{G} \left[(\hat{\mathbf{p}}_{(1g)} - \hat{\mathbf{p}}_{(2g)})(\hat{\mathbf{p}}_{(1g)} - \hat{\mathbf{p}}_{(2g)})' \right] \tag{C.6}$$

where $\hat{\mathbf{p}}_{(1g)}$ is the set of SPREE estimates based on the allocation structure derived from the replicate formed by leaving out one half-sample in the gth stratum but including twice the other selection in that stratum. $\hat{\mathbf{p}}_{(2g)}$ is the set of SPREE estimates based on the allocation structure derived from the complement replicate formed by interchanging the eliminated duplicated selections in the gth stratum. G is the total number of strata.

Generalized Jackknife:

$$V_{D1} = \frac{n-m}{bm} \sum_{k=1}^{b} (\hat{\mathbf{p}}_k - \hat{\mathbf{p}})(\hat{\mathbf{p}}_k - \hat{\mathbf{p}})' \tag{C.7}$$

$$V_{D2} = \frac{n-m}{bm} \sum_{k=1}^{b} (\hat{\mathbf{p}}_k - \hat{\hat{\mathbf{p}}})(\hat{\mathbf{p}}_k - \hat{\hat{\mathbf{p}}})' \tag{C.8}$$

where $\hat{\mathbf{p}}_k$ is the set of SPREE estimates based on the allocation structure derived from the kth sub-sample of size $n-m$, such that n is the total sample size, $b = n/m$ is an integer and $\hat{\hat{\mathbf{p}}} = \frac{1}{b}\sum_{k=1}^{b} \hat{\mathbf{p}}_k$

Table C.1: Variance estimate of survey margins using Linearization and Half Jackknife methods

Variables	Margins	Variance	
		Linearization Method	Jackknife Method
Non-poor	55,206,524	162,398,650,368	162,397,805,010
Poor	23,502,040	95,990,161,408	95,990,291,292
Wall_strong	45,575,436	158,438,998,016	158,437,996,180
Wall_light	18,086,702	78,671,962,112	78,672,137,864
Wall_salvaged	976,596	4,841,352,704	4,841,352,631
Wall_others	14,069,827	103,245,307,904	103,245,148,410
Rural	39,952,672	321,161,953,280	321,163,449,420
Urban	38,755,892	340,814,659,584	340,814,261,120
Female Headed HH	10,724,622	23,549,622,272	23,549,801,609
Male Headed HH	67,983,944	157,199,319,040	157,200,156,110
No HS	19,658,640	48,505,962,496	48,506,401,548
HS	59,049,924	147,846,250,496	147,845,393,680
Roof_strong	52,285,904	188,136,095,744	188,135,238,450
Roof_light	15,349,936	61,414,191,104	61,414,074,162
Roof_salvaged	712,601	3,219,569,408	3,219,567,235
Roof_others	10,360,120	85,407,989,760	85,407,863,976
No domhelp	76,534,352	159,374,049,280	159,374,485,890
With domhelp	2,174,214	6,059,106,816	6,059,105,695
Single	72,365,824	171,552,325,632	171,549,997,300
Duplex	2,581,274	11,691,931,648	11,691,918,039
Apartment/Condo	3,409,318	19,842,443,264	19,842,425,767
Industrial/Agricultural Building	335,851	2,009,924,736	2,009,926,577
Others	16,297	28,312,864	28,312,860
With Educ	73,269,656	158,051,041,280	158,050,956,850
No Educ	5,438,905	18,904,578,048	18,904,625,245
Not Elem	15,585,754	32,653,936,640	32,654,005,855
With Elem	63,122,808	150,818,717,696	150,819,693,520
No Coed	45,207,660	139,399,512,064	139,398,443,790
With Coed	33,500,904	105,756,516,352	105,756,164,420
With Spouse	67,840,832	153,896,337,408	153,897,281,170
No Spouse	10,867,731	20,983,627,776	20,983,350,659

Appendix D

Crosstabulation of Poverty Status with the Auxiliary Variables

Table D.1: Poverty status by marital status cross-tabulation

Poverty Status	2000		2003	
	With Spouse	No Spouse	With Spouse	No Spouse
Non-poor	0.8292	0.1708	0.839	0.161
	0.6393	0.7722	0.6827	0.8179
Poor	0.9028	0.0972	0.9158	0.0842
	0.3607	0.2278	0.3173	0.1821
Total	0.8543	0.1457	0.8619	0.1381
Pearson Chi-square:	386.7387		433.548	

Table D.2: Poverty status by strong wall material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.3188	0.6812	0.3118	0.6882
	0.4785	0.7994	0.5195	0.8337
Poor	0.6702	0.3298	0.6774	0.3226
	0.5215	0.2006	0.4805	0.1663
Total	0.4387	0.5613	0.421	0.579
Pearson Chi-square:	4459.5161		4797.5341	

Table D.3: Poverty status by light wall material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.8708	0.1292	0.8667	0.1333
	0.7504	0.361	0.7893	0.4069
Poor	0.5588	0.4412	0.5435	0.4565
	0.2496	0.639	0.2107	0.5931
Total	0.7643	0.2357	0.7702	0.2298
Pearson Chi-square:	4803.2307		5161.2799	

Table D.4: Poverty status by salvaged wall material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9917 0.6614	0.0083 0.4376	0.9901 0.7032	0.0099 0.5619
Poor	0.9793 0.3386	0.0207 0.5624	0.9818 0.2968	0.0182 0.4381
Total	0.9874	0.0126	0.9876	0.0124
Pearson Chi-square:	109.2187		48.7632	

Table D.5: Poverty status by other wall material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.8188 0.6662	0.1812 0.6266	0.8315 0.7101	0.1685 0.6613
Poor	0.7917 0.3338	0.2083 0.3734	0.7972 0.2899	0.2028 0.3387
Total	0.8095	0.1905	0.8212	0.1788
Pearson Chi-square:	42.4242		69.7406	

Table D.6: Poverty status by strong roof material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.2511 0.4571	0.7489 0.7729	0.24 0.5014	0.76 0.8025
Poor	0.5754 0.5429	0.4246 0.2271	0.5606 0.4986	0.4394 0.1975
Total	0.3618	0.6382	0.3357	0.6643
Pearson Chi-square:	4048.9553		4032.3445	

Table D.7: Poverty status by light roof material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.8922 0.7426	0.1078 0.3402	0.8932 0.7783	0.1068 0.384
Poor	0.5967 0.2574	0.4033 0.6598	0.5977 0.2217	0.4023 0.616
Total	0.7914	0.2086	0.805	0.195
Pearson Chi-square:	4701.4031		4865.9742	

Table D.8: Poverty status by salvaged roof material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9935 0.6608	0.0065 0.4379	0.9921 0.7022	0.0079 0.6093
Poor	0.9839 0.3392	0.0161 0.5621	0.9882 0.2978	0.0118 0.3907
Total	0.9903	0.0097	0.9909	0.0091
Pearson Chi-square:	84.359		15.4579	

Table D.9: Poverty status by other roof material cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.8631 0.6637	0.1369 0.6285	0.8747 0.7065	0.1253 0.6678
Poor	0.8439 0.3363	0.1561 0.3715	0.8536 0.2935	0.1464 0.3322
Total	0.8566	0.1434	0.8684	0.1316
Pearson Chi-square:	26.744		34.1642	

Table D.10: Poverty status by male headed household cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.169	0.831	0.1624	0.8376
	0.7909	0.637	0.8358	0.6802
Poor	0.0862	0.9138	0.0749	0.9251
	0.2091	0.363	0.1642	0.3198
Total	0.1408	0.8592	0.1363	0.8637
Pearson Chi-square:	503.9128		568.3383	

Table D.11: Poverty status by household employs domestic helper cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9545	0.0455	0.9609	0.0391
	0.6482	0.9937	0.6931	0.993
Poor	0.9994	5.60E-04	0.9994	0.00065
	0.3518	0.0063	0.3069	0.007
Total	0.9699	0.0301	0.9724	0.0276
Pearson Chi-square:	613.2583		481.5416	

Table D.12: Poverty status by living in a single type of house cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.0832	0.9168	0.1029	0.8971
	0.8788	0.644	0.8957	0.6844
Poor	0.0221	0.9779	0.0282	0.9718
	0.1212	0.356	0.1043	0.3156
Total	0.0624	0.9376	0.0806	0.9194
Pearson Chi-square:	566.8021		659.6854	

Table D.13: Poverty status by living in a duplex type of house cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9626	0.0374	0.9609	0.0391
	0.6535	0.8232	0.6968	0.8365
Poor	0.9845	0.0155	0.982	0.018
	0.3465	0.1768	0.3032	0.1635
Total	0.9700	0.0300	0.9672	0.0328
Pearson Chi-square:	147.0846		123.3804	

Table D.14: Poverty status by living in a multiple(apartment/condominium) type of house cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9596	0.0404	0.9422	0.0578
	0.6505	0.9351	0.6908	0.9364
Poor	0.9946	0.0054	0.9908	0.0092
	0.3495	0.0649	0.3092	0.0636
Total	0.9716	0.0284	0.9567	0.0433
Pearson Chi-square:	393.3504		498.7101	

Table D.15: Poverty status living in a Commercial/Industrial/Agricultural Building/House type of house cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9948	0.0052	0.9943	0.0057
	0.6577	0.9151	0.7004	0.9412
Poor	0.9991	0.0009	0.9992	0.0008
	0.3423	0.0849	0.2996	0.0588
Total	0.9963	0.0037	0.9957	0.0043
Pearson Chi-square:	43.3054		49.1475	

Table D.16: Poverty status by living in other type of house (cave, boat, etc.) cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9998	0.0002	0.9998	0.0002
	0.6586	0.5802	0.7014	0.8088
Poor	0.9997	0.0003	0.9999	0.0001
	0.3414	0.4198	0.2986	0.1912
Total	0.9998	0.0002	0.9998	0.0002
Pearson Chi-square:	0.2611		0.4764	

Table D.17: Poverty status by urbanity cross-tabulation

Poverty Status	2000		2003	
	Rural	Urban	Rural	Urban
Non-poor	0.406	0.594	0.4127	0.5873
	0.5168	0.8107	0.5702	0.8366
Poor	0.7324	0.2676	0.7306	0.2694
	0.4832	0.1893	0.4298	0.1634
Total	0.5174	0.4826	0.5076	0.4924
Pearson Chi-square:	3790.6335		3536.2094	

Table D.18: Poverty status by having a family member with college education cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.4131	0.5869	0.4551	0.5449
	0.4902	0.8687	0.5557	0.898
Poor	0.8289	0.1711	0.8545	0.1455
	0.5098	0.1313	0.4443	0.102
Total	0.555	0.445	0.5744	0.4256
Pearson Chi-square:	6222.7155		5708.2611	

Table D.19: Poverty status by having a family member (10 years old and over) with no education cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.9463	0.0537	0.9553	0.0447
	0.6799	0.4249	0.7198	0.4534
Poor	0.8597	0.1403	0.8735	0.1265
	0.3201	0.5751	0.2802	0.5466
Total	0.9167	0.0833	0.9309	0.0691
Pearson Chi-square:	872.8756		910.0856	

Table D.20: Poverty status by having a family member (10 years old and over) with only high school education cross-tabulation

Poverty Status	2000		2003	
	No	Yes	No	Yes
Non-poor	0.2158	0.7842	0.2175	0.7825
	0.5768	0.6854	0.6108	0.7316
Poor	0.3055	0.6945	0.3255	0.6745
	0.4232	0.3146	0.3892	0.2684
Total	0.2464	0.7536	0.2498	0.7502
Pearson Chi-square:	385.2571		544.6208	

Appendix E

Correlation Matrix for the Margins

Table E.1: Correlation matrix for the margins

	Npoor	Poor	Wlstr	Wlght	Wlsvg	Wloth	Rural	Urban	FHd	MHd	NHs	HS	Rfstr	Rflgt	Rfsvg	Rfoth	NDh	WDh
Npoor	1.00																	
Poor	-0.38	1.00																
Wlstr	0.56	0.02	1.00															
Wlght	0.05	0.36	-0.15	1.00														
Wlsvg	0.07	0.01	-0.05	-0.07	1.00													
Wloth	0.14	0.15	-0.38	-0.25	0.01	1.00												
Rural	0.16	0.21	0.07	0.27	-0.03	0.07	1.00											
Urban	0.34	0.07	0.33	-0.04	0.08	0.10	-0.75	1.00										
FHd	0.27	-0.03	0.21	0.00	0.01	0.05	-0.02	0.19	1.00									
MHd	0.62	0.41	0.50	0.34	0.07	0.24	0.33	0.32	-0.14	1.00								
NHs	0.10	0.34	0.12	0.31	-0.03	0.05	0.22	0.04	0.05	0.35	1.00							
HS	0.69	0.21	0.53	0.17	0.10	0.23	0.21	0.39	0.23	0.77	-0.19	1.00						
Rfstr	0.58	0.06	0.83	0.02	-0.05	-0.27	0.13	0.30	0.18	0.56	0.14	0.57	1.00					
Rflgt	0.01	0.37	-0.12	0.71	0.03	-0.11	0.22	0.00	0.00	0.30	0.33	0.13	-0.22	1.00				
Rfsvg	0.01	0.02	-0.07	-0.05	0.69	0.02	-0.06	0.08	0.00	0.03	-0.06	0.06	-0.12	-0.02	1.00			
Rfoth	0.11	0.14	-0.33	-0.17	0.02	0.83	0.07	0.08	0.07	0.19	0.03	0.21	-0.42	-0.10	0.04	1.00		
NDh	0.67	0.42	0.53	0.34	0.08	0.26	0.33	0.36	0.23	0.92	0.35	0.83	0.59	0.31	0.03	0.22	1.00	
WDh	0.25	-0.11	0.22	-0.02	-0.01	-0.04	-0.07	0.19	0.12	0.13	0.08	0.13	0.20	-0.02	-0.03	-0.03	-0.02	1.00
Single	0.60	0.40	0.53	0.33	0.10	0.18	0.33	0.31	0.21	0.85	0.35	0.76	0.56	0.30	0.05	0.15	0.90	0.12
Duplex	0.16	-0.04	0.05	0.00	0.03	0.09	-0.04	0.13	0.08	0.09	-0.06	0.16	0.05	0.01	0.04	0.08	0.11	0.08
Aprtmt	0.14	-0.03	0.07	-0.04	-0.12	0.12	-0.04	0.12	0.03	0.11	0.04	0.10	0.08	-0.04	-0.13	0.10	0.11	0.06
IndAgB	0.00	-0.02	-0.03	0.02	0.04	-0.01	0.00	-0.01	0.02	-0.03	0.05	-0.05	-0.04	0.00	0.07	0.01	-0.04	0.08
Others	-0.02	0.01	0.00	-0.02	0.00	0.00	0.01	-0.02	-0.03	0.00	0.00	-0.01	0.00	-0.01	0.02	-0.01	-0.01	0.00
Weduc	0.71	0.31	0.57	0.25	0.07	0.25	0.27	0.39	0.24	0.87	0.26	0.85	0.61	0.22	0.03	0.22	0.92	0.17
Neduc	0.03	0.26	0.04	0.26	0.03	0.00	0.14	0.02	0.03	0.22	0.33	0.05	0.06	0.25	0.01	0.01	0.23	0.00
Nelem	0.44	-0.16	0.30	0.01	-0.01	0.02	-0.07	0.29	0.16	0.26	0.03	0.31	0.29	0.01	-0.05	0.00	0.26	0.28
Welem	0.53	0.48	0.46	0.34	0.08	0.25	0.36	0.27	0.18	0.85	0.36	0.74	0.51	0.31	0.05	0.22	0.90	0.05
Ncoed	0.22	0.56	0.19	0.39	0.04	0.24	0.40	0.07	-0.02	0.68	0.37	0.48	0.25	0.36	-0.01	0.24	0.69	-0.12
Wcoed	0.62	-0.16	0.49	-0.04	0.04	0.04	-0.06	0.41	0.33	0.38	0.03	0.51	0.48	-0.04	0.05	-0.01	0.43	0.35
Wspo	0.63	0.40	0.52	0.34	0.07	0.23	0.32	0.34	-0.02	0.96	0.35	0.79	0.57	0.30	0.02	0.19	0.92	0.15
Nspo	0.26	0.01	0.19	0.01	0.03	0.08	0.03	0.16	0.75	-0.02	0.07	0.24	0.18	0.01	0.02	0.09	0.25	0.08

See codes in Table 7.5

Table E.2: Correlation matrix for the margins - continuation from previous page

	Single	Duplex	Aprtmt	IndAgB	Others	Weduc	Neduc	Nelem	Welem	Ncoed	Wcoed	Wspo	Nspo
Single	1.00												
Duplex	-0.14	1.00											
Aprtmt	-0.20	-0.03	1.00										
IndAgB	-0.08	0.05	-0.19	1.00									
Others	-0.04	-0.01	0.04	-0.01	1.00								
Weduc	0.85	0.11	0.12	-0.03	0.00	1.00							
Neduc	0.21	0.03	-0.01	0.01	-0.04	-0.12	1.00						
Nelem	0.24	0.09	0.12	0.02	-0.03	0.30	0.05	1.00					
Welem	0.84	0.08	0.07	-0.03	0.00	0.84	0.21	-0.14	1.00				
Ncoed	0.61	0.04	0.08	-0.05	0.00	0.56	0.29	-0.11	0.73	1.00			
Wcoed	0.43	0.11	0.05	0.03	-0.01	0.52	-0.06	0.52	0.27	-0.33	1.00		
Wspo	0.85	0.11	0.12	-0.04	-0.01	0.88	0.21	0.27	0.85	0.66	0.40	1.00	
Nspo	0.24	0.06	0.01	0.04	-0.01	0.25	0.05	0.14	0.21	0.03	0.30	-0.10	1.00

See codes Table 7.5

Appendix F

Do Files in Stata for the ESPREE Updating Method

cen_prep1.do

```
//This program will generate the provincial level data containing information
//on the variable of interest and census counts on each municipality.
//The files ending with “s” are created here.
clear set mem 500m cd “E:\Office computer \Maris\Thesis\Cen_Res”
forvalues iR=1/16      {
use “E:\Office computer\Maris\PHfiles\Pnames.dta”, clear
keep if region==‘iR’
local iP=_N
save PNtemp, replace
forvalues ip=1/‘iP’      {
use PNtemp, clear
local Fname=pname[‘ip’]+”v.dta”
local Pname=pname[‘ip’]+”r.dta”
local Sname=pname[‘ip’]+”s.dta”
use “‘Fname’”, clear keep ic bcode regn prov urb famsize wall_* roof_* urb head_male
all_hsed all_hsed dom_help
compress
recode wall_strong (1=1), gen(cwall_strong)
recode wall_light (1=2), gen(cwall_light)
recode wall_salvaged (1=3), gen(cwall_salvaged)
gen wall=cwall_strong+cwall_light+cwall_salvaged
keep if wall< ·
replace wall=4 if wall==0
gen wall_oth=(wall==4)
compress
recode roof_strong (1=1), gen(croof_strong)
recode roof_light (1=2), gen(croof_light)
recode roof_salvaged (1=3), gen(croof_salvaged)
gen roof=croof_strong+croof_light+croof_salvaged
keep if roof< ·
compress
replace roof=4 if roof==0
gen roof_oth=(roof==4)
replace head_male=(head_male> 0)
replace all_hsed=(all_hsed> 0)
replace all_hsed=(all_hsed> 0)
replace dom_help=(dom_help> 0)
compress
sort ic
preserve use “‘Pname’”, clear
sort ic
save “‘Pname’”, replace
restore
merge ic using “‘Pname’”
keep if Yb1< .
keep if _merge==3 drop _merge
gen mcode=int(bcode/1000)
sort prov
merge prov using “E:\ Office computer\ Maris\ Thesis\ PovLines.dta”
drop if _merge==2
gen pline=urbline
replace pline=rurline if urb==0
drop _merge urbline rurline
* Fix up Marawi City (moved province): replace pline=12910 if prov==98 & mun==17
local i=1
while ‘i’<=100      {
replace Yb‘i’=(Yb‘i’<pline)
local i=‘i’+1
}
compress
forval j=1/100      {
preserve
gen freqc‘j’=1
compress
collapse (count) freqc‘j’ [pw=famsize], by(Yb‘j’ wall roof dom_help urb head_male
all_hsed mcode) fast rename Yb‘j’ Yb
compress
if ‘j’==1      {
sort Yb wall roof dom_help urb head_male all_hsed mcode
save “‘Sname’”, replace
}
if ‘j’> 1      {
merge Yb wall roof dom_help urb head_male all_hsed mcode using “‘Sname’”
drop _merge
sort Yb wall roof dom_help urb head_male all_hsed mcode
save “‘Sname’”, replace
}
restore
}
```

```
}
}
```

cen_prep2.do

```
//This program will combine the provincial level data generated from
//the cen_prep1 stata do file

clear
set mem 500m
cd “E:\Office computer\Maris\Thesis\Cen_Res”
forvalues iR=1/16      {
use “E:\Office computer\Maris\PHfiles\Pnames.dta”, clear
keep if region==‘iR’
local iP=_N
save PNtemp, replace
forvalues ip=1/‘iP’      {
use PNtemp, clear
local Sname=pname[‘ip’]+”s.dta”
use ”‘Sname’”, clear
gen regn=int(mcode/10000)
gen prov=int(((mcode/10000)-regn)*100)
compress
if ‘ip’==1      {
save Temp, replace
}
if ‘ip’>1      {
append using Temp
save Temp, replace
}
}
if ‘iR’==1      {
save POmun, replace
}
if ‘iR’>1      {
append using POmun
save POmun, replace
}
}
order regn prov mcode Yb wall urb head_male all_hsed sort regn prov
save “E:\Tempdata\Final\POmun”, replace
//to consider those categories that do not exist in
//the census, we merge the P0mun file to the codes file
//so those categories will be included and will be assigned
//the value zero
```

```
preserve
infile mcode Yb wall urb head_male all_hsed roof dom_help
using “E:\Tempdata\Final\mcodes1.csv”, clear
sort Yb wall urb head_male all_hsed roof dom_help mcode
save “E:\Tempdata\Final\Codes_new”, replace
restore
sort Yb wall urb head_male all_hsed roof dom_help mcode merge Yb wall urb head_male
all_hsed roof dom_help mcode
using “E:\Tempdata\Final\Codes_new”
replace regn=int(mcode/10000) if regn==.
replace prov=int(((mcode/10000)-regn)*100) if prov==.
forval i=1/100       {
replace freqc‘i’=0 if freqc‘i’==.
}
drop _merge
save “E:\Tempdata\Final1\POmun”, replace
order regn prov mcode Yb wall urb head_male all_hsed roof dom_help
save “E:\Tempdata\Final1\POmun”, replace
```

surv_prep1a.do

```
//(A) This program prepares the survey data (combined FIES and LFS)
//to generate the needed replicated survey margins. Since the survey data
//is composed of certainty and non-certainty PSUs, here we first removed the
//the certainty PSUs and generate BRR weights only for the non-certainty PSUs
//a file containing the data of non-certainty PSUs and BRR weights is the output.
clear
set memory 500m
set matsize 2000
cd “E:\Tempdata\Data”
use fieslfs03_nocrtnPSU, clear //this file contains only the non-certainty PSUs sort
prov
/*preparation of the data and the hadamard matrix*/
preserve
sort strata bcode
gen n=1
collapse (count) n, by(strata bcode)
sort strata bcode by strata: gen bcode1=_n
drop n
sort strata bcode
save “E:\Tempdata\Data\strata_bcode1”, replace
restore
sort strata bcode
merge strata bcode using “E:\Tempdata\Data\strata_bcode1”
```

```
keep if _merge==3
drop _merge
sort strata
merge strata using “E:\Tempdata\Data\strata_had100.dta” /*uses a hadamard matrix of size 100*/
keep if _merge==3 drop _merge
sort prov
merge prov using “D:\Maris\Corsairdisk\progs_data(my comp)2\progs_data\PovLines03.dta”
keep if _merge==3
compress
gen pline=urbline
replace pline=rurline if urb==0
drop _merge urbline rurline
gen Yb=(incpp<pline)
drop pline
forval i=1/100    {
gen brr_wt‘i’=2*sswgtpp if h‘i’==1 & bcode1==1
replace brr_wt‘i’=2*sswgtpp if h‘i’==0 & bcode1==2
replace brr_wt‘i’=0 if brr_wt‘i’==.
}
gen brr_wt101=sswgtpp //the last replicate is the full sample (but excluding the
certainty psus)//
drop h1-h100 n bcode1
sort strata bcode
save fieslfs03_brwts, replace
```

surv_prep1b.do

```
//(B) Survey Preparation
// This program will get the margins for the survey data using the full data set
// (certainty and non-certainty PSUs) with the survey weights available from the
FIES and LFS data set.
clear
set memory 500m
cd “E:\”
use “E:\Tempdata\Data\fieslfs03.dta”, clear
  sort prov
merge prov using “D:\Maris\Corsairdisk\progs_data(my comp)2\progs_data\PovLines03.dta”
keep if _merge==3
gen pline=urbline
replace pline=rurline if urb==0
drop _merge urbline rurline
gen Yb=(incpp<pline)
/*recoding of variables*/
```

```
rename stratum strata
gen wall_strong=(wall==1)
gen wall_light=(wall==2)
gen wall_salvaged=(wall==3)
gen wall_oth=(wall==4)
compress
gen roof_strong=(roof==1)
gen roof_light=(roof==2)
gen roof_salvaged=(roof==3)
gen roof_oth=(roof==4)
compress
gen head_male=(head_sex==1)
drop head_sex
gen no_spouse=(head_status!=2)
drop head_status
/*generation of margins*/
gen domain=0
gen freqs=1
svyset bcode [pweight=sswgtpp], strata(strata)
egen index1=group(Yb), label
egen index2=group(wall), label
egen index3=group(urb), label
egen index4=group(head_male), label
egen index5=group(all_hsed), label
egen index6=group(roof), label
egen index7=group(dom_help), label
foreach i of numlist 1/2     {
replace domain=(index1==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
foreach i of numlist 1/4     {
replace domain=(index2==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
foreach i of numlist 1/2       {
replace domain=(index3==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
foreach i of numlist 1/2       {
replace domain=(index4==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
foreach i of numlist 1/2       {
replace domain=(index5==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
foreach i of numlist 1/4       {
replace domain=(index6==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
foreach i of numlist 1/2       {
```

```
replace domain=(index7==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A=(‘i’,t,s)
matrix B=nullmat(B)\A
}
svmat B rename B1 id2 rename B2 count
format count %20.6f
rename B3 var
format var%20.6f
keep id2 count
keep if id2< .
rename count count102 /*since there are 100 brr replicates, the 101th is the full
sample with no certain psu, so make the 101st obs the full sample*/
gen ic=_n
sort ic
save “E:\Tempdata\Final1\margins_allnew”, replace
```

surv_prep1c.do

```
//(C) Survey Preparation
// This program will get the margins for the survey data using the BRR weights
// and only the non-certainty PSUs. All the steps are similar to (B), however
// the data set used here is the one generated from (A)
clear
set memory 500m
cd “E:\Tempdata\Final1”
use “E:\Tempdata\Data\fieslfs03_brwts.dta”, clear // file from (A)
sort prov
/*recoding of the variables*/
gen wall_strong=(wall==1)
gen wall_light=(wall==2)
gen wall_salvaged=(wall==3)
gen wall_oth=(wall==4)
compress
gen roof_strong=(roof==1)
gen roof_light=(roof==2)
gen roof_salvaged=(roof==3)
gen roof_oth=(roof==4)
compress
gen head_male=(head_sex==1)
```

```
drop head_sex
gen no_spouse=(head_status!=2)
drop head_status
/*generation of margins*/
gen domain=0
gen freqs=1
egen index1=group(Yb), label
egen index2=group(wall), label
egen index3=group(urb), label
egen index4=group(head_male), label
egen index5=group(all_hsed), label
egen index6=group(roof), label
egen index7=group(dom_help), label
forval j= 1/101      {
svyset bcode [pweight=brr_wt‘j’], strata(strata)
foreach i of numlist 1/2      { replace domain=(index1==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
foreach i of numlist 1/4      {
replace domain=(index2==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
foreach i of numlist 1/2      {
replace domain=(index3==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
foreach i of numlist 1/2      {
replace domain=(index4==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
foreach i of numlist 1/2      { replace domain=(index5==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
foreach i of numlist 1/4      {
replace domain=(index6==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
foreach i of numlist 1/2      {
replace domain=(index7==‘i’)
svy: total freqs, subpop(domain)
matrix T=e(b)
matrix V=e(V)
scalar t=T[1,1]
scalar s=V[1,1]
matrix A‘j’=(‘i’,t,s)
matrix B‘j’=nullmat(B‘j’)\A‘j’
}
svmat B‘j’
compress
rename B‘j’1 id2‘j’
rename B‘j’2 count‘j’
drop B‘j’3
format count‘j’ %20.6f
drop brr_wt‘j’
}
drop index*
keep id21 count* order id21 count*
keep if id2< .
gen ic=_n sort ic save survmrgnwu_ncPSU, replace
merge ic using margins_allnew order ic id21 id2
drop _merge id21 drop ic id2
save surv_margins1, replace
```

≫ **Run R program for generation of Pseudo-counts**

R codes for generating the pseudo-counts from the survey data.

```
library(foreign) Margins < −read.dta(“H:/Rprogram/BRR/surv_margins91_102.dta”)
Y < − as.matrix(Margins)
Yb < − Y[1:2,]
Wall< − Y[3:6,]
Urb < − Y[7:8,]
HM < − Y[9:10,]
AH < − Y[11:12,]
Roof< −Y[13:16,]
Dom< −Y[17:18,]
A< −array(0, dim=c(1623,12,2,4,2,2,2,4,2))
for (m in 1:1623)
   {
     for (l in 1:12)
       {
       for (i in 1:2)
         {
         for (j in 1:4)
           {
           for (k in 1:2)
             {
             for (n in 1:2)
               {
               for (o in 1:2)
                 {
                 for (p in 1:4)
                   {
                   for (q in 1:2)
                     {
```

```
A[m,l,i,j,k,n,o,p,q]=(Yb[i,l]*Wall[j,l]*Urb[k,l]*HM[n,l]*AH[o,l]*Roof[p,l]*Dom[q,l]/(sum(Yb[,l])ˆ6))/1623
                      }
                    }
                  }
                }
              }
            }
          }
        }
      }
b< − NULL
for (i in 1:2)
    {
    for (k in 1:2)
        {
        for (n in 1:2)
          {
          for (o in 1:2)
            {
            for (q in 1:2)
                {
              for ( p in 1:4)
                {
                for (j in 1:4)
                  {
                  for (m in 1:1623)
                    {
          b < − rbind(b, A[m,,i,j,k,n,o,p,q]) }
                  }
                }
              }
            }
          }
        }
      }
data < − read.fwf(‘H:/Rprogram/mcodefile.txt’,widths=c(10),col.names=c(‘code’))
    k < − matrix(0,nrow=1623*512,ncol=8)
    for (i in 1:4) {
    k[((i-1)*1623+1):(i*1623),1]< − data$code
    k[((i-1)*1623+1):(i*1623),2] < − 0
    k[((i-1)*1623+1):(i*1623),3] < − i
    k[((i-1)*1623+1):(i*1623),4] < − 0
    k[((i-1)*1623+1):(i*1623),5] < − 0
    k[((i-1)*1623+1):(i*1623),6] < − 0
    k[((i-1)*1623+1):(i*1623),7] < − 1
    k[((i-1)*1623+1):(i*1623),8] < − 0
    }
    for (i in 5:8) {
    k[((i-1)*1623+1):(i*1623),1] < − data$code
    k[((i-1)*1623+1):(i*1623),2] < − 0
    k[((i-1)*1623+1):(i*1623),3] < −i-4
    k[((i-1)*1623+1):(i*1623),4] < − 0
    k[((i-1)*1623+1):(i*1623),5] < − 0
    k[((i-1)*1623+1):(i*1623),6] < − 0
    k[((i-1)*1623+1):(i*1623),7] < − 2
    k[((i-1)*1623+1):(i*1623),8] < − 0
    }
    for (i in 9:12) {
    k[((i-1)*1623+1):(i*1623),1] < − data$code
    k[((i-1)*1623+1):(i*1623),2] < − 0
    k[((i-1)*1623+1):(i*1623),3] < − i-8
    k[((i-1)*1623+1):(i*1623),4] < − 0
    k[((i-1)*1623+1):(i*1623),5] < − 0
    k[((i-1)*1623+1):(i*1623),6] < − 0
    k[((i-1)*1623+1):(i*1623),7] < − 3
    k[((i-1)*1623+1):(i*1623),8] < − 0
    }
                    •
                    •
                    •
    for (i in 505:508) {
    k[((i-1)*1623+1):(i*1623),1] < − data$code
    k[((i-1)*1623+1):(i*1623),2] < − 1
    k[((i-1)*1623+1):(i*1623),3] < − i-504
    k[((i-1)*1623+1):(i*1623),4] < − 1
    k[((i-1)*1623+1):(i*1623),5] < − 1
    k[((i-1)*1623+1):(i*1623),6] < − 1
    k[((i-1)*1623+1):(i*1623),7] < − 3
    k[((i-1)*1623+1):(i*1623),8] < − 1
    }
    for (i in 509:512) {
    k[((i-1)*1623+1):(i*1623),1] < − data$code
    k[((i-1)*1623+1):(i*1623),2] < − 1
    k[((i-1)*1623+1):(i*1623),3] < − i-508
    k[((i-1)*1623+1):(i*1623),4] < − 1
    k[((i-1)*1623+1):(i*1623),5] < − 1
    k[((i-1)*1623+1):(i*1623),6] < − 1
    k[((i-1)*1623+1):(i*1623),7] < − 4
    k[((i-1)*1623+1):(i*1623),8] < − 1
    }
marg< −cbind(k,b)
write(t(marg),‘H:/Rprogram/BRR/Pseudocounts91_102.csv’,ncolumns=20)
```

infiling_all.do

// This program is used to convert and combine the data file generated from the R program (.csv files) into Stata data file.

```
clear
set mem 800m
preserve
infile mcode Yb wall urb head_male all_hsed roof dom_help PS1-PS10
using “H:\Rprogram\BRR\Pseudocounts1_10.csv”, clear
sort mcode Yb wall urb head_male all_hsed roof dom_help
save “H:\Rprogram\BRR\psdcounts1_10”, replace
restore
preserve
```

```
infile mcode Yb wall urb head_male all_hsed roof dom_help PS11-PS20
using “H:\Rprogram\BRR\Pseudocounts11_20.csv”, clear
sort mcode Yb wall urb head_male all_hsed roof dom_help
save “H:\Rprogram\BRR\psdcount11_20”, replace
restore
use “H:\Rprogram\BRR\psdcounts1_10”
sort mcode Yb wall urb head_male all_hsed roof dom_help
merge mcode Yb wall urb head_male all_hsed roof dom_help
using “H:\Rprogram\BRR\psdcount11_20”
drop _merge
sort mcode Yb wall urb head_male all_hsed roof dom_help
save “H:\Rprogram\BRR\psdcount_new3”, replace
compress
                    •
                    •
                    •
preserve
infile mcode Yb wall urb head_male all_hsed roof dom_help PS91-PS102
using “H:\Rprogram\BRR\Pseudocounts91_102.csv”, clear
sort mcode Yb wall urb head_male all_hsed roof dom_help
save “H:\Rprogram\BRR\psdcount91_102”, replace
restore
merge mcode Yb wall urb head_male all_hsed roof dom_help
using “H:\Rprogram\BRR\psdcount91_102”
drop _merge
sort mcode Yb wall urb head_male all_hsed roof dom_help
save “H:\Rprogram\BRR\psdcount_new3”, replace
compress


//The following programs will fit the loglinear model and generate estimates of
//poverty incidence using the bootstrap estimates of the census data and the
//BRR estimates of survey
```

gen_povinc1a.do

```
clear
set mem 1G
set matsize 5000
use “E:\Tempdata\Final1\POmun” //is the file of the census counts
sort Yb wall urb head_male all_hsed mcode
compress
save “E:\Tempdata\Final1\POmun”,
replace
preserve
use “H:\Rprogram\BRR\psdcount_new3” //the file of the margins from BRR
sort Yb wall urb head_male all_hsed mcode
save “H:\Rprogram\BRR\psdcount_new3”, replace
compress
restore
merge Yb wall urb head_male all_hsed mcode
using “H:\Rprogram\BRR\psdcount_new3”
keep if _merge==3
drop _merge
```

```
compress
forval i=1/100      {
replace freqc‘i’=freqc‘i’+0.00001
compress
}
gen wall_strong=(wall==1)
gen wall_light=(wall==2)
gen wall_salvaged=(wall==3)
compress
drop PS1-PS100
//to generate 100 bootstrap estimates using the pseudocensus data and the original/complete survey
data
forval j=1/100      {
glm PS101 Yb wall_strong wall_light wall_salvaged urb head_male all_hsed, family(poisson) link(log)
lnoffset(freqc‘j’)
predict botpred‘j’, mu
compress
drop freqc‘j’
}
keep regn prov mcode Yb wall wall_strong wall_light wall_salvaged urb head_male all_hsed botpred*
compress
sort regn prov mcode Yb wall wall_strong wall_light wall_salvaged urb head_male all_hsed
save “E:\Tempdata\Final1\BCounts”, replace
```

gen_povinc1b.do

```
clear
set mem 1G
set matsize 5000
use “E:\Tempdata\Final1\POmun” //is the file of the census counts
sort Yb wall urb head_male all_hsed mcode
compress
save “E:\Tempdata\Final1\POmun” , replace
preserve
use “H:\Rprogram\BRR\psdcount_new3” //the file of the margins from BRR
sort Yb wall urb head_male all_hsed mcode
compress
save “H:\Rprogram\BRR\psdcount_new3”, replace
restore
merge Yb wall urb head_male all_hsed mcode
using “H:\Rprogram\BRR\psdcount_new3”
keep if _merge==3
drop _merge
compress
forval i=1/100 {
replace freqc‘i’=freqc‘i’+0.00001
compress
}
egen Meancount=rmean(freqc100-freqc1)
compress
gen wall_strong=(wall==1)
gen wall_light=(wall==2)
gen wall_salvaged=(wall==3)
```

```
compress
drop freqc100-freqc1
//to generate the brr estimate using the average of the pseudocensus data and the brr estimates of
the survey data
forval i=1/102 {
glm PS‘i’ Yb wall_strong wall_light wall_salvaged urb head_male all_hsed, family(poisson) link(log)
lnoffset(Meancount) predict jkpred‘i’, mu
compress
drop PS‘i’
}
keep regn prov mcode Yb wall wall_strong wall_light wall_salvaged urb head_male all_hsed size1
jkpred*
compress
sort regn prov mcode Yb wall wall_strong wall_light wall_salvaged urb head_male all_hsed
save “E:\Tempdata\Final1\JCounts”, replace
merge regn prov mcode Yb wall wall_strong wall_light wall_salvaged urb head_male all_hsed
using “E:\Tempdata\Final1\BCounts”
keep if _merge==3
save “E:\Tempdata\Final1\BJCounts”, replace
```

gen_povinc2.do

```
clear
set mem 1G
use “E:\Tempdata\Final1\BJCounts” //is the file of the census counts
keep regn prov mcode Yb botpred*
preserve
collapse (sum) botpred1-botpred100, by (mcode Yb) fast
compress
global botpredvars ”botpred*”
foreach x of varlist $botpredvars {
egen ‘x’p=total(‘x’), by(mcode) //p for municipal total//
compress
gen st‘x’=‘x’/‘x’p //st for status//
compress
drop ‘x’ ‘x’p
}
drop if Yb==0
order mcode stbotpred*
egen sdbot=rowsd(stbotpred1-stbotpred100)
compress
gen varbot=sdbotˆ2
compress
keep mcode varbot
sort mcode
save “E:\Tempdata\Final1\botvar_mun”, replace
restore
```

gen_povinc3.do

```
clear
set mem 1G
use “E:\Tempdata\Final1\BJCounts” //is the file of the census counts
```

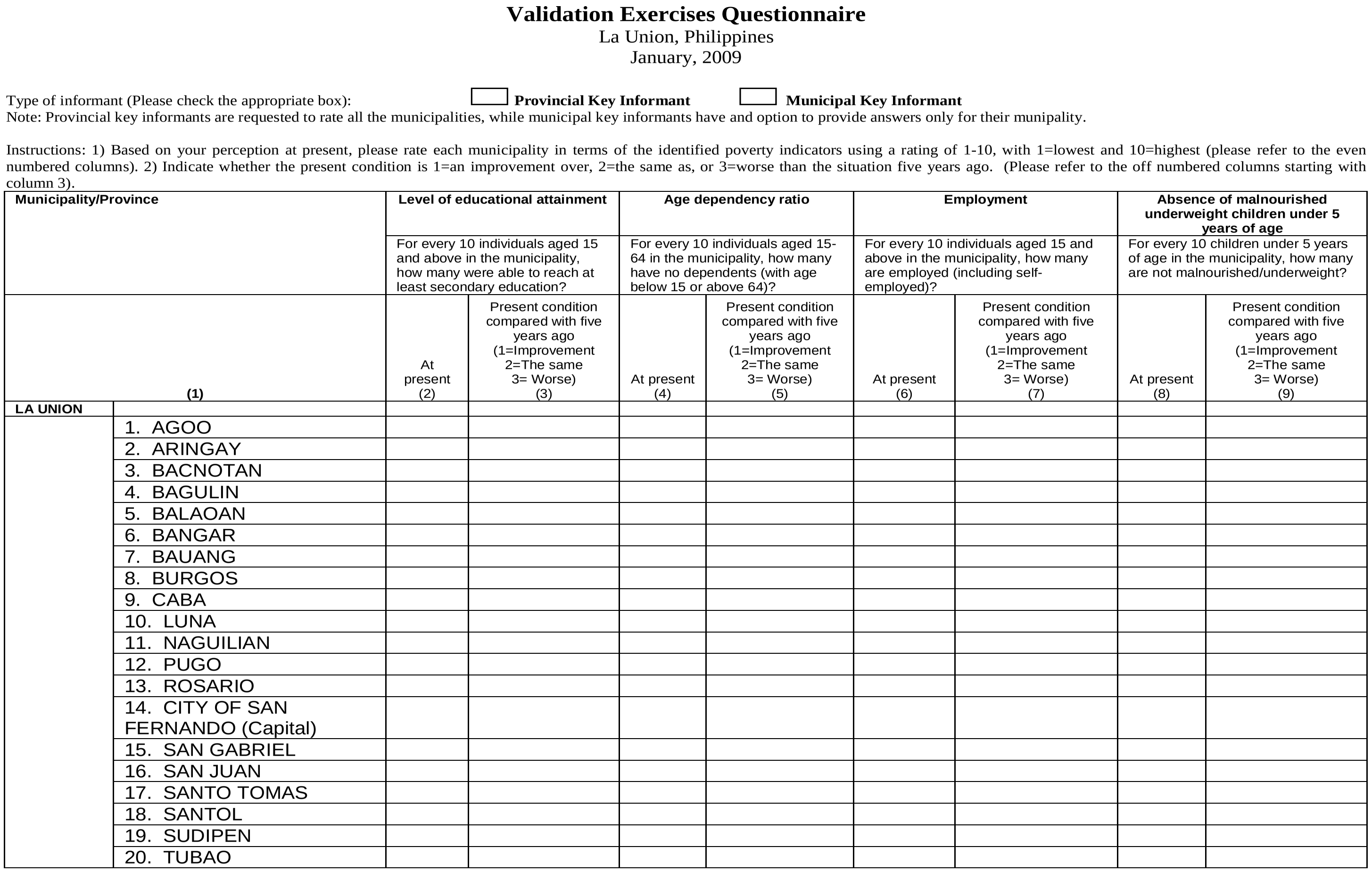
```
keep regn prov size1 mcode Yb jkpred*
preserve
collapse (mean) regn prov (sum) jkpred1-jkpred102, by (mcode Yb) fast
compress
global jkpredvars ”jkpred*”
foreach x of varlist $jkpredvars {
egen ‘x’p=total(‘x’), by(mcode)
compress
gen st‘x’=‘x’/‘x’p
compress
drop ‘x’ ‘x’p
}
drop if Yb==0
gen sttjkpred=stjkpred101
compress forval i=1/100 {
gen dif2_‘i’=(stjkpred‘i’-sttjkpred)*(stjkpred‘i’-sttjkpred)
compress
drop stjkpred‘i’
}
order regn prov mcode dif2_*
egen RSumsq=rowtotal(dif2_1-dif2_100)
compress
drop dif2_*
format RSumsq %20.10f
gen Vbrr=(1/100)*RSumsq
compress
format Vbrr %20.10f
gen SDbrr=sqrt(Vbrr)
compress
compress keep regn prov mcode Vbrr stjkpred102
keep if mcode< .
save “E:\Tempdata\Final1\br_estvar_mun”, replace
sort mcode
merge mcode using “E:\Tempdata\Final1\botvar_mun.dta”
keep if _merge==3
gen SE=sqrt(Vbrr+varbot)
drop _merge
order regn prov mcode sort regn prov mcode save
“E:\Tempdata\Final1\povincid_mun”, replace
restore
preserve
collapse (mean) regn (sum) jkpred1-jkpred102, by (prov Yb) fast
compress
global jkpredvars1 ”jkpred*”
foreach x of varlist $jkpredvars1 {
egen ‘x’p=total(‘x’), by(prov)
compress
gen st‘x’=‘x’/‘x’p
compress
drop ‘x’ ‘x’p
}
drop if Yb==0
gen sttjkpred=stjkpred101
```

```
compress
forval i=1/100 {
gen dif2_‘i’=(stjkpred‘i’-sttjkpred)*(stjkpred‘i’-sttjkpred)
compress
drop stjkpred‘i’
}
order regn prov dif2_*
egen RSumsq=rowtotal(dif2_1-dif2_100)
compress
drop dif2_*
format RSumsq %20.10f
gen Vbrr=(1/100)*RSumsq
compress
format Vbrr %20.10f
gen SDbrr=sqrt(Vbrr)
compress
compress
keep regn prov Vbrr stjkpred102
keep if prov< .
save “E:\Tempdata\Final1\br_estvar_prov”, replace
sort prov
merge prov using “E:\Tempdata\Final1\botvar_prov.dta”
keep if _merge==3
gen SE=sqrt(Vbrr+varbot)
drop _merge
save “E:\Tempdata\Final1\povincid_prov”, replace
restore
preserve
collapse (sum) jkpred1-jkpred102, by (regn Yb) fast
compress
global jkpredvars2 ”jkpred*”
foreach x of varlist $jkpredvars2 {
egen ‘x’p=total(‘x’), by(regn)
compress
gen st‘x’=‘x’/‘x’p
compress
drop ‘x’ ‘x’p
}
drop if Yb==0
gen sttjkpred=stjkpred101
compress
forval i=1/100 {
gen dif2_‘i’=(stjkpred‘i’-sttjkpred)*(stjkpred‘i’-sttjkpred)
compress
drop stjkpred‘i’
}
order regn dif2_*
egen RSumsq=rowtotal(dif2_1-dif2_100)
compress
drop dif2_*
format RSumsq %20.10f
gen Vbrr=(1/100)*RSumsq
compress
```

```
format Vbrr %20.10f
gen SDbrr=sqrt(Vbrr)
compress
compress
keep regn Vbrr stjkpred102
keep if regn< .
save “E:\Tempdata\Final1\br_estvar_regn”, replace
sort regn
merge regn using “E:\Tempdata\Final1\botvar_regn.dta”
keep if _merge==3
gen SE=sqrt(Vbrr+varbot)
drop _merge
save “E:\Tempdata\Final1\povincid_regn”, replace
restore
```

Appendix G

Sample Validation Form



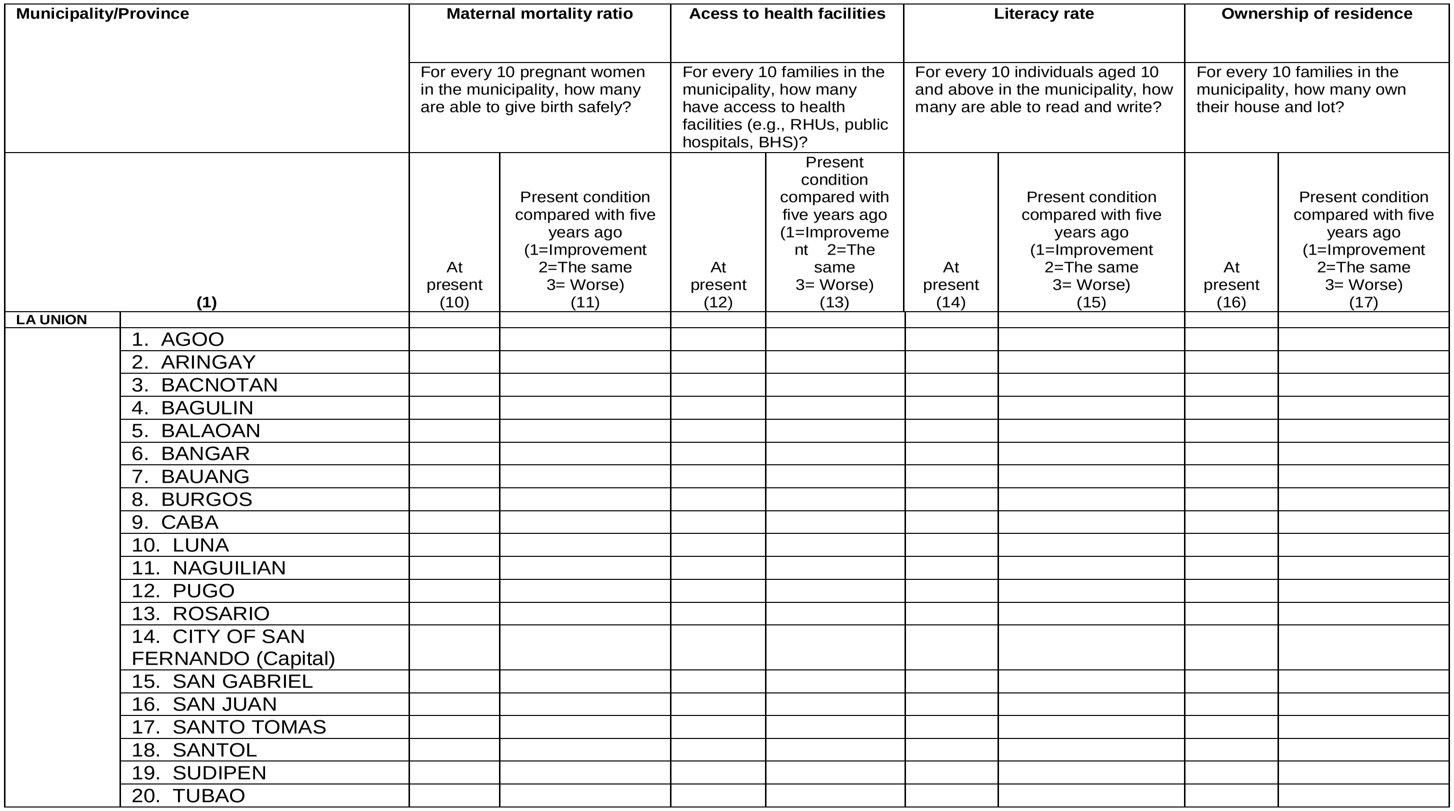
Validation Exercises Questionnaire
La Union, Philippines
January, 2009

Type of informant (Please check the appropriate box): □ **Provincial Key Informant** □ **Municipal Key Informant**
Note: Provincial key informants are requested to rate all the municipalities, while municipal key informants have and option to provide answers only for their munipality.

Instructions: 1) Based on your perception at present, please rate each municipality in terms of the identified poverty indicators using a rating of 1-10, with 1=lowest and 10=highest (please refer to the even numbered columns). 2) Indicate whether the present condition is 1=an improvement over, 2=the same as, or 3=worse than the situation five years ago. (Please refer to the off numbered columns starting with column 3).

Municipality/Province		**Level of educational attainment**		**Age dependency ratio**		**Employment**		**Absence of malnourished underweight children under 5 years of age**	
		For every 10 individuals aged 15 and above in the municipality, how many were able to reach at least secondary education?		For every 10 individuals aged 15-64 in the municipality, how many have no dependents (with age below 15 or above 64)?		For every 10 individuals aged 15 and above in the municipality, how many are employed (including self-employed)?		For every 10 children under 5 years of age in the municipality, how many are not malnourished/underweight?	
	(1)	At present (2)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (3)	At present (4)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (5)	At present (6)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (7)	At present (8)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (9)
LA UNION									
	1. AGOO								
	2. ARINGAY								
	3. BACNOTAN								
	4. BAGULIN								
	5. BALAOAN								
	6. BANGAR								
	7. BAUANG								
	8. BURGOS								
	9. CABA								
	10. LUNA								
	11. NAGUILIAN								
	12. PUGO								
	13. ROSARIO								
	14. CITY OF SAN FERNANDO (Capital)								
	15. SAN GABRIEL								
	16. SAN JUAN								
	17. SANTO TOMAS								
	18. SANTOL								
	19. SUDIPEN								
	20. TUBAO								

Figure G.1: Validation Form Part 1

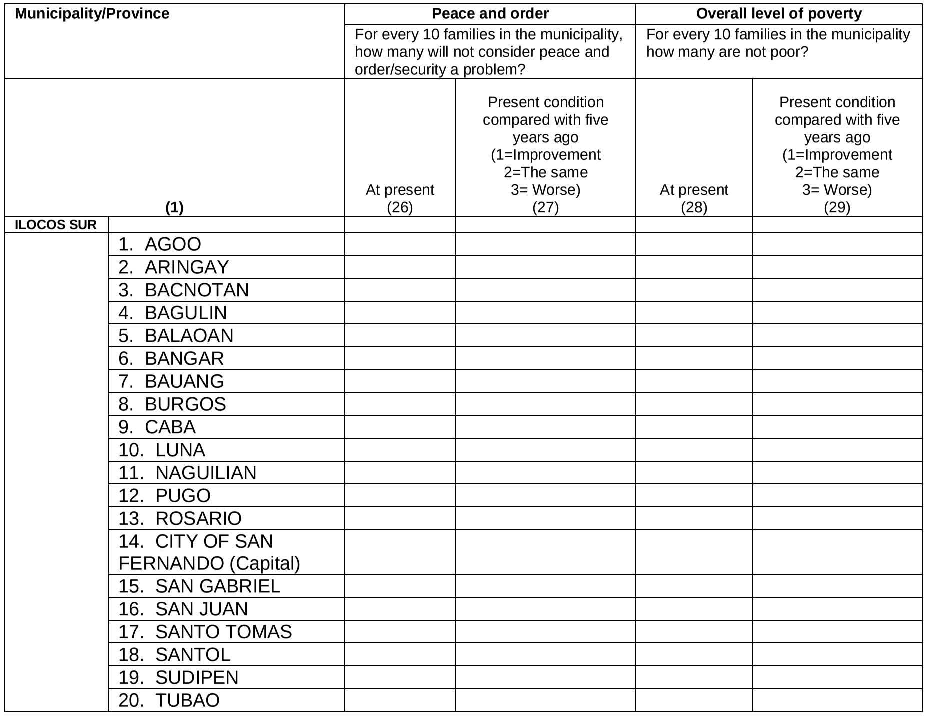


Municipality/Province		**Maternal mortality ratio**		**Acess to health facilities**		**Literacy rate**		**Ownership of residence**	
		For every 10 pregnant women in the municipality, how many are able to give birth safely?		For every 10 families in the municipality, how many have access to health facilities (e.g., RHUs, public hospitals, BHS)?		For every 10 individuals aged 10 and above in the municipality, how many are able to read and write?		For every 10 families in the municipality, how many own their house and lot?	
	(1)	At present (10)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (11)	At present (12)	Present condition compared with five years ago (1=Improveme nt 2=The same 3= Worse) (13)	At present (14)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (15)	At present (16)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (17)
LA UNION									
	1. AGOO								
	2. ARINGAY								
	3. BACNOTAN								
	4. BAGULIN								
	5. BALAOAN								
	6. BANGAR								
	7. BAUANG								
	8. BURGOS								
	9. CABA								
	10. LUNA								
	11. NAGUILIAN								
	12. PUGO								
	13. ROSARIO								
	14. CITY OF SAN FERNANDO (Capital)								
	15. SAN GABRIEL								
	16. SAN JUAN								
	17. SANTO TOMAS								
	18. SANTOL								
	19. SUDIPEN								
	20. TUBAO								

Figure G.2: Validation Form Part 2

Municipality/Province		**Quality of housing**		**Acess to safe water**		**Access to sanitary toilet**		**Access to electricity**	
		For every 10 families in the municipality, how many have houses made of strong construction materials (galvanized iron/aluminum, tile, concrete, brick stone or asbestos)?		For every 10 families in the municipality, how many have access to safe water (faucet, tubed or piped well)?		For every 10 families in the municipality, how many have acess to sanitary toilets (water-sealed or closed pit type)?		For every 10 families in the municipality, how many have access to electricity?	
	(1)	At present (18)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (19)	At present (20)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (21)	At present (22)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (23)	At present (24)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (25)
LA UNION									
	1. AGOO								
	2. ARINGAY								
	3. BACNOTAN								
	4. BAGULIN								
	5. BALAOAN								
	6. BANGAR								
	7. BAUANG								
	8. BURGOS								
	9. CABA								
	10. LUNA								
	11. NAGUILIAN								
	12. PUGO								
	13. ROSARIO								
	14. CITY OF SAN FERNANDO (Capital)								
	15. SAN GABRIEL								
	16. SAN JUAN								
	17. SANTO TOMAS								
	18. SANTOL								
	19. SUDIPEN								
	20. TUBAO								

Figure G.3: Validation Form Part 3



Municipality/Province		Peace and order		Overall level of poverty	
		For every 10 families in the municipality, how many will not consider peace and order/security a problem?		For every 10 families in the municipality how many are not poor?	
	(1)	At present (26)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (27)	At present (28)	Present condition compared with five years ago (1=Improvement 2=The same 3= Worse) (29)
ILOCOS SUR					
	1. AGOO				
	2. ARINGAY				
	3. BACNOTAN				
	4. BAGULIN				
	5. BALAOAN				
	6. BANGAR				
	7. BAUANG				
	8. BURGOS				
	9. CABA				
	10. LUNA				
	11. NAGUILIAN				
	12. PUGO				
	13. ROSARIO				
	14. CITY OF SAN FERNANDO (Capital)				
	15. SAN GABRIEL				
	16. SAN JUAN				
	17. SANTO TOMAS				
	18. SANTOL				
	19. SUDIPEN				
	20. TUBAO				

Figure G.4: Validation Form Part 4