CIMAT

Centro de Investigación en Matemáticas, A.C.

STATISTICAL MODELS FOR SMALL AREA ESTIMATION

TESIS

Que para obtener el Grado de:

Maestro en Ciencias con Orientación en Probabilidad y Estadística

PRESENTA:

Georges Bucyibaruta

Director:

Dr. Rogelio Ramos Quiroga

Guanajuato, Gto, México, Noviembre de 2014

Integrantes del Jurado.

Presidente: Dr. Miguel Nakamura Savoy (CIMAT).

Secretario: Dr. Enrique Raúl Villa Diharce (CIMAT).

Vocal y Director de la Tesis: Dr. Rogelio Ramos Quiroga (CIMAT).

Asesor:

Dr. Rogelio Ramos Quiroga

Sustentante:

Georges Bucyibaruta

Abstract

The demand for reliable small area estimates derived from survey data has increased greatly in recent years due to, among other things, their growing use in formulating policies and programs, allocation of government funds, regional planning, small area business decisions and other applications.

Traditional area-specific (direct) estimates may not provide acceptable precision for small areas because sample sizes are rarely large enough in many small areas of interest. This makes it necessary to borrow information across related areas through indirect estimation based on models, using auxiliary information such as recent census data and current administrative data. Methods based on models are now widely accepted. Popular techniques for small area estimation use explicit statistical models, to indirectly estimate the small area parameters of interest.

In this thesis, a brief review of the theory of Linear and Generalized Linear Mixed Models is given, the point estimation focusing on Restricted Maximum Likelihood. Bootstrap methods for two-Level models are used for the construction of confidence intervals. Model-based approaches for small area estimation are discussed under a finite population framework. For responses in the Exponential family we discuss the use of the Monte Carlo Expectation and Maximization algorithm to deal with problematic marginal likelihoods.

Dedication

This Thesis is dedicated
To
God Almighty,
my mother,
my sisters and brothers,
and the memory of my father.

Acknowledgements

Throughout the period of my study, I have had cause to be grateful for the advice, support and understanding of many people. In particular I would like to express my boundless gratitude to my supervisor Dr. Rogelio Ramos Quiroga for helping me get into the problem of small area estimation and kindly direct this work. He has guided me in my research through the use of his profound knowledge, insightful thinking, thoughtful suggestions as well as his unflagging enthusiasm. I really appreciate the enormous amount of time and patience he has spent on me.
This work has also benefited greatly from the advice, comments and observations of Dr. Miguel Nakamura Savoy and Dr. Enrique Raúl Villa Diharce.

As well as the people mentioned above, my special thanks go to members of department of probability statistics of CIMAT for the interesting lectures they offered to us and their consistent support and encouragement throughout my study. A mention must also go to my wonderful classmates, current and former students and the CIMAT community in general who provided pleasant atmosphere and useful distraction during my stay in Guanajuato.

I gratefully acknowledge the admission to CIMAT and financial support from Secretaria de Relaciones Exteriores (SRE), México with which it has been possible to realize this achievement.

Finally but certainly not least, I want to remember my entire family in Rwanda whose unlimited support, prayers, and encouragement carried me through times of despondency and threatening despair.

My sincere thanks to all.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Small Area Estimation (Rao, 2003) tackles the important statistical problem of providing reliable estimates of a variable of interest in a set of small geographical areas. The main problem is that it is almost impossible to measure the value of the target variable for all the individuals in the areas of interest and, hence, a survey is conducted to obtain a representative sample.

Surveys are usually designed to provide reliable estimates of finite population parameters of interest for large domains in which the large domains can refer to geographical regions such as states, counties, municipalities or demographic groups such as age, race, or gender. The information collected is used to produce some direct estimate of the target variable that relies only on the survey design and the sampled data. Unfortunately, sampling from all areas can be expensive in resources and time. Domain estimators that are computed using only the sample data from the domain are known as direct estimators. Even if direct estimators have several desirable design-based properties, direct estimates often lack precision when domain sample sizes are small. In some cases there might be a need for estimates of domains which are not considered during planning or design, and this leads to the possibility that the sample size of the domains could be small and even zero. Direct estimators of small area parameters may have large variance and sometimes cannot be calculated due to lack of observations (Rao, 2003,Chapiter 1, Montanari, Ranalli and Vicarelli, 2010).

To produce estimates for small areas with an adequate level of precision often requires indirect estimators that use auxiliary data through a linking model in order to borrow strength and increase the effective sample size. Regression models are often used in this context to provide estimates for the non-sampled areas. The model is fitted on data from the sample and used together with additional information from auxiliary data to compute estimates for non-sampled areas. This method is often known as synthetic estimation

(Gonzlez, 1973).
Small area estimation (SAE) involves the class of techniques to estimate the parameters of interest from a subpopulation that is considered to be small. In this context, a small area can be seen as a small geographical region or a demographic group which has a sample size that is too small for delivering direct estimates of adequate precision (Rao, 2003, Chapiter 1). The overriding question of small area estimation is how to obtain reliable estimates when the sample data contain too few observations (even zero in some cases) for statistical inference of adequate precision due to the fact that the budget and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas as we mentioned in the previous paragraph.

1.1 Background

The use of small area statistical methods has been in existence dating back to as early as the eleventh and seventeenth centuries in England and Canada respectively. In these cases, they were based on either census or on administrative records (Brackstone, 1987). In recent years, the statistical techniques for small area estimation have been a focus for various authors in the context of sample surveys, and there is an ever-growing demand for reliable estimates of small area populations of all types (both public and private sectors). Some of the main reasons of that demand that we can pick out are: Administrative planning, disease mapping, determination of state funding allocations, etc. Examples of small area estimation applications include: The prediction of the area under corn and soybeans in counties of Iowa (Battese et al., 1988), the production of estimates of income and poverty measures for various population subgroups for states, counties and school districts (Rao, 2003, Chapiter 1).

Traditional methods of indirect estimation mentioned provide a link or a connection to related small areas through supplementary data and make specific allowance for between area variation. In particular, the mixed model involving random area-specific effects that account for between area variation.

Mixed models with unit level auxiliary data have been used for small area estimation by a good number of authors. Battese, Harter and Fuller (1988) use a linear mixed model to predict the area planted with corn and soybeans in Iowa counties. In this case, the area under corn and soybeans are considered to be small because the sample sizes obtained were too small for direct estimation. Datta and Ghosh (1991) introduce the hierarchical Bayes predictor for general mixed models. Larsen (2003) compared estimators for proportions based on two unit level models, a single model with no area level model covariates and a model using the area level information. Malec (2005) proposes Bayesian small area estimates for means of binary responses using a multivariate binomial (multinomial) model. Jiang

(2007) reviews the classical inferential approach for linear and generalized linear mixed models and discusses the prediction for a function of fixed and random effects. Montanari, Ranalli, and Vicarelli (2010) consider unit level linear mixed models and logistic mixed models, for binary response variables and fully known auxiliary information. Vizcaino, Cortina, Morales Gonzalez (2011) derive small area estimators for labor force indicators in Galicia using a multinomial logit mixed model.

1.2 Structure of the Thesis

In Section 1.1 we introduced the idea of small area estimation and highlighted some concepts of small area estimation applied to a wide variety of situations. The work presented in this thesis can be divided into two broad categories, the first dealing with a linear mixed model approach with emphasis on small area estimation. The second category involves the use of generalized linear mixed models in small area estimation. The rest of this thesis is structured in the following manner. Chapter 2 gives an overview of approaches to small area estimation. In Chapter 3, the general theory on linear mixed models is introduced, two approaches for parameter estimation are considered: Point estimation and confidence interval estimation. Furthermore, the theory of prediction is applied to find the Empirical Best Linear Unbiased Predictor (EBLUP) of the random effects. In Chapter 4, the theory on linear mixed models and the related EBLUP theory revised in Chapter 3 is applied to the estimation of small area quantities of interest using the nested error regression model based on the general theorem of prediction. In Chapter 5, the class of generalized linear mixed models is introduced and its various aspects such as model formulation, estimation and the Monte Carlo Expectation Maximization (EM) algorithm are presented. Chapter 6 looks at the Small Area Estimation under mixed effects logistic models. The application and contextualization of the generalized linear mixed models and the related Empirical Best Predictor (EBP) is implemented in the context of small area estimation. Finally, general conclusions and suggestions for future research are presented in Chapter 7.

1.3 Main Contribution of the Thesis

The overall objective of this work is to review the various statistical models and then consequently implement their applications in the small area estimation context. In doing the review, we considered the Linear Mixed Models (LMM) and the Generalized Linear Mixed models (GLMM). Firstly, we give a review on theory of the Linear Mixed Model and applications to small area estimation under the normality assumption using the MLE and REML methods, and we further explain the derivation of the Best Linear Unbiased Predictor /Empirical the Best Linear Unbiased Predictor. Secondly, we contextualize the Generalized Linear Mixed Models approach in small area estimation with emphasis on making inferences on the model parameters when applied to count data. Given the structure of these models, we show that the direct ML and REML methods are limited due to the complexity of the (marginal) likelihood function and we suggest the Monte Carlo EM as an alternative method of parameter estimation that avoids of the complexities of direct ML methods.

Chapter 2

Approaches to Small Area Estimation

2.1 Introduction

In this chapter, we briefly describe some approaches to small area estimation. In general we can differentiate two types of small area estimators: direct and indirect estimators. The direct small area estimator is based on survey design, while indirect estimators are mainly based on different models and techniques such as implicit model based approaches that include synthetic and composite estimators known as traditional indirect estimators, and explicit models which are categorized as area level and unit level models and are based on mixed model methodologies (Rao, 2003, Chapiter 5).

In this chapter we will briefly talk about, direct small area estimators, implicit models (synthetic and composite estimators) and explicitly small area models (unit level model and area level model).

2.2 Direct Estimation

Following the definition given by Rao (2003), a small area estimator is direct when it uses the sample values of study variables from the specified area only. In general, the direct estimators are unbiased estimators, however, due to the small sample size, such estimators might have unacceptably large standard errors. Within this type of estimators there is the expansion estimator. Assume that the quantity of interest is the total Y and there is no auxiliary information. The expansion estimator of Y is

$$\hat{Y} = \sum_{s} w_i y_i,$$

where w_i are the design weights, s is a selected sample, $i \in s$.

2.3 Indirect Estimation

Indirect or model-based small area estimators rely on statistical models to provide estimates for all small areas. Once the model is chosen, its parameters are estimated using the data obtained in the survey. An important issue in indirect small area estimation is that auxiliary information or covariates are needed.

2.3.1 Synthetic Estimation

The synthetic estimator is an example of an estimator, which can be considered either model-based or design-based model-assisted. In both cases the specified linear relationship between y (study variable) and the auxiliary variables, described with the parameter β (vector of regression coefficients) plays an important role. In the design-based approach we assume no explicit model, but the more correlated y is with the auxiliaries the more efficient is the estimator. The name synthetic estimator comes from the fact that these estimators borrow strength by synthesizing data from many different areas.

It is called a synthetic estimator, the one which is a reliable direct estimator for a large area, covering various small areas, and is used as an estimator for a small area considering that the small area has the same characteristics as the larger area (Gonzlez, 1973, Rao, 2003, Chapiter 4).

2.3.2 Composite Estimation

As we mentioned, as the sample size in a small area increases, a direct estimator becomes more desirable than a synthetic estimator. This means, when area level sample sizes are relatively small the synthetic estimator outperforms the traditional direct estimator. Synthetic estimators have a big influence of information from the other areas, thus they may have small variance but a large bias in the case where the hypothesis of homogeneity is not satisfied.

According to Rao (2003), to avoid the potential bias of a synthetic estimator, say $\hat{Y}_{is}$ and the instability of the direct estimator, say $\hat{Y}_{id}$, we consider a convex combination of both, known as the composite estimator.

$$\hat{Y}_{ic} = \omega_i \hat{Y}_{is} + (1 - \omega_i)\hat{Y}_{id},$$

for a suitable chosen weight ω_i $(0 \leq \omega_i \leq 1)$, where c, s, and d stand for composite, synthetic and direct, respectively.

2.4 Small Area Models

2.4.1 Introduction

Traditional methods of indirect estimators, mentioned above, are based on implicit models (synthetic and composite). We now turn to explicit linking models which provide significant improvements in techniques for indirect estimation. Based on mixed model methodology, these techniques incorporate random effects into the model. The random effects account for the between-area variation that cannot be explained by including auxiliary variables. Most small area models can be defined as an area-level model or a unit-level model (Rao, 2003, Chapiter 5).

2.4.2 Basic Area Level Model

The area level model relates the small area information on the response variable to area-specific auxiliary variables. One of the most widely used area level models for small area estimation was proposed by Fay and Herriot (1979). According to the Fay-Herriot model, a basic area level model assumes that the small area parameter of interest η_i is related to area-specific auxiliary data x_i through a linear model

$$\eta_i = x_i^t\beta + \upsilon_i,\ i = 1, 2, \ldots, m, \tag{2.1}$$

where m is the number of small areas, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^t$ is $p \times 1$ vector of regression coefficients, and the υ_i's are area-specific random effects assumed to be independent and identically distributed (*iid*) with $\mathrm{E}(\upsilon_i) = 0$ and $\mathrm{Var}(\upsilon_i) = \sigma_\upsilon^2$, model expectation and model variance respectively. Normality of the random effects υ_i is also often used, but it is possible to make robust inferences by relaxing the normality assumption (Rao, 2003).

The area level model assumes that there exists a direct survey estimator $\hat{\eta}_i$ for the small area parameter η_i such that

$$\hat{\eta}_i = \eta_i + \epsilon_i,\ i = 1, 2, \ldots, m, \tag{2.2}$$

where the ϵ_i is the sample error associated with the direct estimator $\hat{\eta}_i$, with the assumptions that the ϵ_i's are independent normal random variables with mean $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \tau_i$. Combining these two equations, yields the area level linear mixed model

$$\hat{\eta}_i = x_i^t\beta + \upsilon_i + \epsilon_i. \tag{2.3}$$

2.4.3 Basic Unit Level Model

Unit level models relate the unit values of the study variable to unit-specific auxiliary variables. These variables are related to the unit level values of response through a linear mixed model known as nested error linear regression model. This type of model can be represented by the following equation

$$y_{ij} = x_{ij}^t \beta + \nu_i + \epsilon_{ij}, \tag{2.4}$$

where y_{ij} is the response of unit j, $j = 1, 2, \ldots, n_i$, in area i, $i = 1, 2, \ldots, m$, x_{ij} is the vector of auxiliary variables, β is the vector of regression parameters, ν_i is the random effect of area i and ϵ_{ij} is the individual unit error term. The area effects ν_i are assumed independent with mean zero and variance σ_ν^2. The errors ϵ_{ij} are independent with mean zero and variance σ_ϵ^2. In addition, the ν_i and ϵ_{ij}'s are assumed to be independent.

The nested error unit level regression model (2.4) was first used to model county crop areas in North-Central Iowa, USA (Battese et al., 1988).

2.5 Comparisons Between Model-based with Design-based Approaches

In this Chapter, we have looked at some of the design-based and model-based approaches for small area estimation that have been described in the literature. However, in practice the direct design-based estimators have been highly appreciated, because they are approximately design unbiased. The traditional indirect estimators, such as synthetic and composite estimators, are based on implicit linking models. Synthetic estimators for small areas are derived from direct estimators for a large area that covers several small areas under the assumption that the small areas have the same characteristics as the large area. Composite estimators are basically weighted averages of direct estimators and synthetic estimators. Both synthetic and composite estimators can yield estimates that provide higher precision compared to direct estimators. However, both types of estimators share a common tendency to be design-biased, and the design bias does not necessarily decrease as the sample size increases.

The explicit linking models provide significant improvements in techniques for indirect estimation. Based on mixed model methodology, these techniques incorporate random effects into the model. The random effects account for the between-area variation that cannot be explained by including auxiliary variables. Most small area models can be defined as an area-level model, or a unit-level model. Area-level models relate small-area direct estimators to area-specific auxiliary data. Unit-level models relate the unit values of a study variable to unit-specific auxiliary data .

In small area estimation, the focus is on bias and variance. For a Small sample size, the unbiasedness of the direct estimators, may be of no practical value due to its large variance.

Model-based estimators may be sometimes biased, but they have the advantage of small variances compared to the design-based estimators (Rao, 2003, Chapiter 5).

Chapter 3

Linear Mixed Models (LMMs)

3.1 Introduction

In this chapter we present an overview of the linear mixed model. It is not intended to be an all encompassing exposition on the subject; the rationale is to briefly explore the methods for parameter estimation used throughout the thesis. In particular, it serves as an introduction to the models that will be used in Chapter 4. The model-based small area estimation largely employs linear mixed models involving random area effects. The auxiliary variables are introduced in the fixed part of the model as covariates.

A linear mixed model (LMM) can be viewed as a parametric linear model for clustered, or longitudinal data that defines the relationships between a continuous dependent variable and various predictor variables (covariates) (Brady et al., 2007). The term mixed refers to the presence of fixed and random effects. To moderate the random variation in the dependent variable at different levels of the data, random effects are directly used since they are associated with one or more random factors.

Generally, clustered data are defined as data sets obtained after measuring the variable of interest on different individuals grouped in a cluster. Longitudinal data, are viewed as data sets in which the dependent variable (observed variable) is measured several times over time for each individual.

LMMs can be considered as hierarchical models with at least two levels of data. Therefore, clustered, and longitudinal data are referred to as hierarchical data sets, since the observations can be placed into levels of a hierarchy in the data. Considering clustered, or longitudinal data sets as multilevel data sets with at least two-level, we have the following: Level 1 denotes observations at the most detailed level of the data. In a clustered data set, Level 1 represents the units of analysis (or subjects) in the study. In a longitudinal data set, Level 1 represents the repeated measures made on the same unit of analysis. Level 2 represents the next level of the hierarchy. In clustered data sets, Level 2 observations represent clusters of units. For longitudinal data sets, Level 2 represents the units of analysis (Brady et al., 2007, McCulloch and Searle, 2001, Kubokawa, 2009).

This chapter will cover the following topics, linear mixed model formulation, point estimation of parameters, confidence intervals for fixed parameters, bootstrapping two-Level models and prediction of random effects.

3.2 Model Formulation

There exist different ways to represent and classify LMM which can be expressed at individual level (Level 1) or Level 2, as well as in matrix notation. The common assumption is that of normality for all random effects; in this case the LMM is called Gaussian Mixed Models (Brady et al., 2007, Jiang, 2007).

In this section we present some alternative representations of LMMs, such are a general matrix specification, hierarchical linear model and marginal linear model which help to write a explicit expression of likelihood function.

3.2.1 General Specification of Model

According to Brady et al. (2007), a simple and general form that indicates how most of the components of an LMM can be written at the level of an individual observation (Level 1) in the context of a clustered two-level data set. By convention, the index j refers Level 1 units and index i means Level 2 units.

$$\begin{aligned} y_{ij} = \underbrace{x_{ij}^t\beta}_{\text{fixed}} + \underbrace{z_{ij}^t u_i}_{\text{random}} + \underbrace{e_{ij}}_{\text{random}} \quad , \\ i = 1, \dots, m; \\ j = 1, \dots, n_i \end{aligned} \tag{3.1}$$

where:

$$\begin{aligned}
y_{ij} &= \text{response of } j^{th} \text{ member of cluster } i \\
m &= \text{number of clusters} \\
n_i &= \text{size of cluster } i \\
x_{ij} &= \text{covariate vector of } j^{th} \text{ member of cluster } i \text{ for fixed effects, } \in \mathbb{R}^p \\
\beta &= \text{fixed effects parameter, } \in \mathbb{R}^p \\
z_{ij} &= \text{covariate vector of } j^{th} \text{ member of cluster } i \text{ for random effects, } \in \mathbb{R}^q \\
u_i &= \text{random effect parameter, } \in \mathbb{R}^q,
\end{aligned}$$

with assumptions that

$$u_i \sim \mathbb{N}_q(0, D), \qquad D \in \mathbb{R}^{q\times q}$$

$$e_i = \begin{pmatrix} e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix} \sim \mathbb{N}_{n_i}(0, \Sigma_i), \qquad \Sigma_i \in \mathbb{R}^{n_i \times n_i}$$

$u_1, \ldots, u_m, e_1, \ldots, e_m$ are assumed independent,Σ_i is covariance matrix of error vector e_i in cluster i and D is covariance matrix of random effects u_i.

Elements along the main diagonal of the D matrix represent the variances of each random effect in u_i, and the off-diagonal elements represent the covariances between two corresponding random effects. Since there are q random effects in the model associated with the i^{th} cluster, D is a $q \times q$ matrix that is symmetric and positive definite, and is defined as follows:

$$D = \mathrm{Var}(u_i) := \begin{pmatrix} \mathrm{Var}(u_{i1}) & \mathrm{Cov}(u_{i1}, u_{i2}) & \ldots & \mathrm{Cov}(u_{i1}, u_{iq}) \\ \mathrm{Cov}(u_{i1}, u_{i2}) & \mathrm{Var}(u_{i2}) & \ldots & \mathrm{Cov}(u_{i2}, u_{iq}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(u_{i1}, u_{iq}) & \mathrm{Cov}(u_{i2}, u_{iq}) & \ldots & \mathrm{Var}(u_{iq}) \end{pmatrix}.$$

The variance and covariance components of the D matrix can be treated as parameters stored in a vector denoted by θ_D with elements and dimension depending on the structure of the matrix D.

In contrast to the standard linear model, the residuals associated with clustered observations within the same cluster in an LMM can be correlated. We assume that the n_i residuals in the vector e_i for a given cluster, i, are random variables that follow a multivariate normal distribution with a mean vector 0 and a positive definite symmetric covariance matrix Σ_i which is defined in the following general form:

$$\Sigma_i = \mathrm{Var}(\mathbf{e}_i) := \begin{pmatrix} \mathrm{Var}(e_{i1}) & \mathrm{Cov}(e_{i1}, e_{i2}) & \dots & \mathrm{Cov}(e_{i1}, e_{in_i}) \\ \mathrm{Cov}(e_{i2}, e_{i1}) & \mathrm{Var}(e_{i2}) & \dots & \mathrm{Cov}(e_{i2}, e_{in_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(e_{in_i}, e_{i1}) & \mathrm{Cov}(e_{in_i}, e_{i2}) & \dots & \mathrm{Var}(e_{in_i}) \end{pmatrix}.$$

In the same way as the case of matrix D, the variance and covariance components of the Σ_i matrix are stored in a vector denoted by θ_Σ.

Considering both cases we define the vector, θ that will be used in subsequent sections, which combines all covariance parameters contained in the vectors θ_D and θ_Σ. Often the parameters in θ are simply variance and covariance components, but sometimes they may have other interpretations such as the autoregressive parameter and moving average component which are used in time series settings (e.g. Box and Jenkins, 1970) or the sill and range in geostatistical applications (e.g. Cressie, 1991).

Covariance Structures for the D Matrix

A D matrix is defined to be unstructured if there are no additional constraints on the values of its elements (aside from positive definiteness and symmetry). The symmetry in the $q \times q$ matrix D indicates that the θ_D vector has $q \times (q+1)/2$ parameters. As example of an unstructured D matrix, in the case of an LMM having two random effects associated with the i^{th} cluster, we consider the following structure:

$$D = \mathrm{Var}(u_i) = \begin{pmatrix} \sigma^2_{u_1} & \sigma_{u_1,u_2} \\ \sigma_{u_1,u_2} & \sigma^2_{u_2} \end{pmatrix}.$$

In this case, the vector θ_D contains three covariance parameters:

$$\theta_D = \begin{pmatrix} \sigma^2_{u_1} \\ \sigma_{u_1,u_2} \\ \sigma^2_{u_2} \end{pmatrix}.$$

Taking into account certain constraints on the structure of D, the very commonly used structure is the variance components (or diagonal) structure, in which each random effect in u_i has its own variance, and all covariances in D are defined to be zero. In general, the θ_D vector for the variance components structure requires q covariance parameters, defining the variances on the diagonal of the D matrix. For example, in an LMM having two random effects associated with the i^{th} subject, a variance component D matrix has the following form:

$$D = \mathrm{Var}(u_i) = \begin{pmatrix} \sigma^2_{u_1} & 0 \\ 0 & \sigma^2_{u_2} \end{pmatrix}.$$

In this case, the vector θ_D contains two parameters:

$$\theta_D = \begin{pmatrix} \sigma^2_{u_1} \\ \sigma^2_{u_2} \end{pmatrix}.$$

Covariance Structures for the Σ_i Matrix

The simplest covariance matrix structure for Σ_i is the diagonal one, in which the residuals associated with observations in the same cluster are assumed to be uncorrelated and to have equal variance. The diagonal Σ_i matrix for each cluster i has the following structure:

$$\Sigma_i = \mathrm{Var}(e_i) := \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

Here we see that the diagonal structure implies one parameter in θ_{Σ_i}, which defines the constant variance at each time point:

$$\theta_{\Sigma_i} = \sigma^2.$$

There exist also other structure of covariance matrix Σ_i called compound symmetry structure. Its general form for each subject i is as follows:

$$\Sigma_i = \mathrm{Var}(e_i) := \begin{pmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \dots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \dots & \sigma^2 + \sigma_1^2 \end{pmatrix}.$$

In this case there are two parameters in the θ_{Σ_i} vector that define the variances and covariances in the Σ_i matrix:

$$\theta_{\Sigma_i} = \begin{pmatrix} \sigma^2 \\ \sigma_1 \end{pmatrix}.$$

The other structure which is commonly used as the covariance structure for the Σ_i matrix, is a first-order autoregressive structure and is denoted by $AR(1)$ for which the general form is :

$$\Sigma_i = \mathrm{Var}(e_i) := \left(\begin{array}{cccc} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \dots & \sigma^2 \end{array} \right).$$

The $AR(1)$ structure has only two parameters in the θ vector that define all the variances and covariances in the Σ_i matrix, a variance parameter ($\sigma^2 > 0$) and a correlation parameter ($-1 \leq \rho \leq 1$),

$$\theta_{\Sigma_i} = \left(\begin{array}{c} \sigma^2 \\ \rho \end{array} \right).$$

Note that the $AR(1)$ structure is frequently used to fit models when data sets are characterized to be equally spaced longitudinal observations on the same units of analysis. According to this structure, the observations which are closer to each other in time exhibit higher correlation than observations farther apart in time. In any given analysis, the structure for the Σ matrix that seems to be most appropriate, is determined due to the given observed data and knowledge about the relationships between observations on an individual subject or cluster.

3.2.2 General Matrix Specification

We now consider the general matrix specification of an LMM for a given cluster i,

$$Y_i = X_i\beta + Z_i u_i + e_i \quad i = 1, \dots, m. \tag{3.2}$$

Where

$$X_i := \left(\begin{array}{c} x_{i1}^t \\ \vdots \\ x_{in_i}^t \end{array} \right) \in \mathbb{R}^{n_i \times p},\ Z_i := \left(\begin{array}{c} z_{i1}^t \\ \vdots \\ z_{in_i}^t \end{array} \right) \in \mathbb{R}^{n_i \times q},\ Y_i := \left(\begin{array}{c} y_{i1} \\ \vdots \\ y_{in_i} \end{array} \right) \in \mathbb{R}^{n_i}.$$

The matrix formulation for a LMM is of the form

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} \in \mathbb{R}^n,\ X := \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} \in \mathbb{R}^{n\times p},\ \beta \in \mathbb{R}^p,\ G := \begin{pmatrix} D & & \\ & \ddots & \\ & & D \end{pmatrix} \in \mathbb{R}^{mq\times mq},$$

$$Z := \begin{pmatrix} Z_1 & 0_{n_1\times q} & \dots & 0_{n_1\times q} \\ 0_{n_2\times q} & Z_2 & & \\ \vdots & & \ddots & \\ 0_{n_m\times q} & & & Z_m \end{pmatrix} \in \mathbb{R}^{n\times mq},\quad 0_{n_i\times q} := \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{n_i\times q},$$

$$u := \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} \in \mathbb{R}^{mq},\quad e := \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix} \in \mathbb{R}^n,\quad R := \begin{pmatrix} \Sigma_1 & & 0 \\ & \ddots & \\ 0 & & \Sigma_m \end{pmatrix} \in \mathbb{R}^{n\times n},\ n = \sum_{i=1}^{m} n_i$$

with this, the Linear Mixed Model eq.(3.2) can be rewritten as

$$Y = X\beta + Zu + e, \tag{3.3}$$

where

$$\begin{pmatrix} u \\ e \end{pmatrix} \sim \mathbf{N}_{mq+n}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G & 0_{mq\times n} \\ 0_{n\times mq} & R \end{pmatrix}\right).$$

3.2.3 Hierarchical Linear Model (HLM) Specification of the LMM

It is often convenient to express an LMM in terms of an explicitly defined hierarchy of simpler models, which correspond to the levels of a clustered or longitudinal data set. When LMMs are specified in such a way, they are often referred to as hierarchical linear models (HLMs), or multilevel models (MLMs). The HLM form of an LMM is equivalent to the general LMM introduced in (3.2), and may be implemented for any LMM. Here we present a general form for the HLM specification of LMMs as a two level hierarchical model.

$$\begin{aligned} Y|u &\sim \mathbf{N}_n\left(X\beta + Zu, R\right) \\ u &\sim \mathbf{N}_{mq}\left(0, G\right). \end{aligned} \tag{3.4}$$

3.2.4 Marginal Linear Model

In subsection 3.2.1 we defined a general form for LMM where the random effects are specified to explain the between-cluster variation. LMMs are referred to as cluster-specific models since it is possible for them to perform the cluster-specific inference.

The marginal linear model is different from the LMM due to the fact of having random effects or not. Marginal models do not allow the cluster-specific effects since the random effects are not specified. The following is the marginal model specification implied by an LMM.

The marginal model in matrix form is

$$Y = X\beta + \varepsilon^*, \tag{3.5}$$

where

$$\varepsilon^* := Zu + e = \underbrace{\begin{pmatrix} Z & I_{n\times n} \end{pmatrix}}_{A} \begin{pmatrix} u \\ e \end{pmatrix}$$

with

$$\varepsilon^* \sim \mathbf{N}_n\left(0, V\right)$$

where

$$\begin{aligned} V = A \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} A^t &= \begin{pmatrix} Z & I_{n\times n} \end{pmatrix} \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} Z^t \\ I_{n\times n} \end{pmatrix} \\ &= \begin{pmatrix} ZG & R \end{pmatrix} \begin{pmatrix} Z^t \\ I_{n\times n} \end{pmatrix} \\ &= ZGZ^t + R \\ &= \begin{pmatrix} Z_1 D Z_1 + \Sigma_1 & & 0 \\ & \ddots & \\ 0 & & Z_m D Z_m + \Sigma_m \end{pmatrix} \\ &= \begin{pmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_m \end{pmatrix}. \end{aligned}$$

The implied marginal model defines the marginal distribution of the Y vector:

$$Y \sim \mathbf{N}_n\left(X\beta, V\right).$$

The implied marginal model has an advantage due to the flexibility of carrying out the estimation of fixed model parameters (fixed-effects and variance components) in an LMM without worrying about the random effects.

3.3 Estimation in Linear Mixed Models

In the LMM, we estimate the fixed-effect parameters, β , and the covariance parameters, (*i.e.*, θ_D and θ_{Σ_i} for the D and Σ_i matrices, respectively). In this section, we discuss maximum likelihood (ML) and restricted maximum likelihood (REML) estimation, which are methods commonly used to estimate these parameters.

3.3.1 Maximum Likelihood (ML) Estimation

In general, maximum likelihood (ML) estimation is a method of obtaining estimates of unknown parameters by optimizing a likelihood function. To apply ML estimation, we first construct the likelihood as a function of the parameters in the specified model, based on distributional assumptions. The maximum likelihood estimates (MLEs) of the parameters are the values of the arguments that maximize the likelihood function (*i.e.*, the values of the parameters that make the observed values of the dependent variable most likely, given the distributional assumptions).

In the context of the LMM, we construct the likelihood function of β and θ by referring to the marginal distribution of the dependent variable Y defined in (3.5). The corresponding multivariate normal probability density function, $f(Y|\beta,\theta)$, is :

$$f\,(Y|\beta,\theta) = (2\pi)^{-n/2}|V|^{-1/2}\exp\{-\frac{1}{2}(Y - X\beta)^t V^{-1}(Y - X\beta)\}. \tag{3.6}$$

Recall that the elements of the V matrix are functions of the covariance parameters in θ. So that the log likelihood is

$$l\,(\beta,\theta) = -\frac{1}{2}(Y - X\beta)^t V^{-1}(Y - X\beta) - \frac{1}{2}\ln(|V|) - \frac{n}{2}\ln(2\pi). \tag{3.7}$$

Case 1: Assume θ is Known

We consider a special case of ML estimation for LMMs, in which we assume that θ, and as a result the matrix V, is known. In this case, the only parameters that we estimate are the fixed effects, β. The log-likelihood function, $l(\beta,\theta)$, thus becomes a function of β only. Then we differentiate respect to β:

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= -2X^tV^{-1}(Y - X\beta)\\ &= -2X^tV^{-1}Y + 2X^tV^{-1}\beta\end{aligned}$$

setting to zero, we get

$$\begin{aligned}-2X^tV^{-1}Y + 2X^tV^{-1}X\beta = 0\\ X^tV^{-1}Y = X^tV^{-1}X\beta\end{aligned}$$

We solve that equation, we get the maximum likelihood estimator

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y. \tag{3.8}$$

Which is also known as the Generalized Least Squares (GLS) estimator when it is assumed that Y has a normal distribution.

Case 2: Assume θ is not Known

In this case the previous relation (3.8) becomes

$$\tilde{\beta}(\theta) = (X^t V(\theta)^{-1} X)^{-1} X^t V(\theta)^{-1} Y. \tag{3.9}$$

To estimate $V(\theta)$, we put (3.9) into (3.7) to obtain the profile log likelihood

$$l_p(\theta) = l\left(\tilde{\beta}(\theta), \theta\right) = -\frac{1}{2}(Y - X\tilde{\beta}(\theta))^t V(\theta)^{-1}(Y - X\tilde{\beta}(\theta)) - \frac{1}{2}\ln(|V(\theta)|) - \frac{n}{2}\ln(2\pi)$$

The profile log likelihood is then maximized to find θ_{ML}.

Harville (1977) asserts that, in estimation of variance parameters, however, the ML method suffers from some well known problems. One of them, it tends to give variance estimates that are biased. For the ordinary linear model where $m = 1$, $V = \sigma^2 I$ and $q = 0$, the ML estimator of the single variance component σ^2 has expectation

$$\begin{aligned} E(\hat{\sigma}^2_{ML}) &= E\left\{(Y - X\hat{\beta})^t (Y - X\hat{\beta})/n\right\} \\ &= \frac{1}{n} E\left\{Y^t (I_n - X(X^t X)^{-1} X^t) Y\right\} \\ &= \frac{1}{n}\sigma^2 tr(I_n - X(X^t X)^{-1} X^t) + \frac{1}{n}\beta^t X^t (I_n - X(X^t X)^{-1} X^t) X\beta \\ &= \frac{n - r}{n}\sigma^2 \end{aligned}$$

where r is the rank of matrix X. So the bias is $\sigma^2 r/n$, that can be significant in the case that the number of degrees of freedom $n - r$ is very small.

At least to some extent, these problems of ML estimation can be avoided by adopting another likelihood-based procedure called residual ML or restricted ML (REML).

3.3.2 Restricted Maximum Likelihood (REML) Estimation

The REML method is based on the likelihood principle and has the same merits, like consistency, efficiency and asymptotic normality, as the ML method. The REML method is now a widely preferred approach to estimate variance parameters in mixed models (Searle et al., 1992, Pinheiro and Bates, 2000, Verbeke and Molenberghs, 2000, Diggle et al., 2002). The REML estimates of θ are based on optimization of the REML log-likelihood function. It means we first remove the effect of the fixed variables: Remember that the residuals are uncorrelated with all the fixed variables in the model. The distribution of the residuals is also normal. But the distribution of the residuals no longer depends on the estimates of the fixed effects; it only depends on the variance components. Harville (1977) considered a joint likelihood function of (β, θ), where by integrating the joint likelihood, he calculated the marginal likelihood function depending only on θ (restricted likelihood):

$$l_R(\theta) = \ln\left(\int L(\beta, \theta)\, d\beta\right) \tag{3.10}$$

where

$$\int L(\beta, \theta)\, d\beta = \int |\mathbf{V}(\theta)|^{-\frac{1}{2}} (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}(Y - X\beta)^t V(\theta)^{-1}(Y - X\beta)\right\} d\beta. \tag{3.11}$$

Now consider

$$(Y - X\beta)^t V(\theta)^{-1}(Y - X\beta) = \beta^t \underbrace{X^t V(\theta)^{-1} X}_{A(\theta)} \beta - 2Y^t V(\theta)^{-1} X\beta + Y^t V(\theta)^{-1} Y$$

Define:

$$B(\theta) = A(\theta)^{-1} X^t V(\theta)^{-1}.$$

Adding and subtracting $Y^t B(\theta)^t A(\theta) B(\theta) Y$, we get

$$\begin{aligned}(Y - X\beta)^t V(\theta)^{-1}(Y - X\beta) = & \ (\beta - B(\theta)Y)^t A(\theta)(\beta - B(\theta)Y) + \\ & Y^t V(\theta)^{-1} Y - Y^t B(\theta)^t A(\theta) B(\theta) Y.\end{aligned}$$

Note that

$$B(\theta)^t A(\theta) = V(\theta)^{-1} X A(\theta)^{-1} A(\theta) = V(\theta)^{-1} X.$$

Therefore we have

$$\begin{aligned}\int L(\beta, \theta)\, d\beta = & \ |\mathbf{V}(\theta)|^{-\frac{1}{2}} (2\pi)^{-n/2} \exp\{-1/2 \times (Y^t [V(\theta)^{-1} + B(\theta)^t A(\theta) B(\theta)] Y\} \times \\ & \int \exp\{-\frac{1}{2}(\beta - B(\theta)Y)^t A(\theta)(\beta - B(\theta)Y)\} d\beta \\ = & \ |\mathbf{V}(\theta)|^{-\frac{1}{2}} (2\pi)^{-n/2} \exp\{-\frac{1}{2}(Y^t [V(\theta)^{-1} + B(\theta)^t A(\theta) B(\theta)] Y\} \frac{(2\pi)^{p/2}}{|A(\theta)^{-1}|^{-1/2}}.\end{aligned}$$

Now

$$\begin{aligned}(Y-X\tilde{\beta}(\theta))^tV(\theta)^{-1}(Y-X\tilde{\beta}(\theta)) = {} & Y^tV(\theta)^{-1}Y-2Y^tV(\theta)^{-1}X\tilde{\beta}(\theta)+\tilde{\beta}(\theta)^t\underbrace{X^tV(\theta)^{-1}X}_{A(\theta)}\tilde{\beta}(\theta)\\ = {} & Y^tV(\theta)^{-1}Y-2Y^tV(\theta)^{-1}X\tilde{\beta}(\theta)+\tilde{\beta}(\theta)^tA(\theta)\tilde{\beta}(\theta)\\ = {} & Y^tV(\theta)^{-1}Y-2Y^tB(\theta)^tA(\theta)B(\theta)Y+Y^tB(\theta)^tA(\theta)B(\theta)Y\\ = {} & Y^tV(\theta)^{-1}Y-Y^tB(\theta)^tA(\theta)B(\theta)Y.\end{aligned}$$

Here we used:

$$\tilde{\beta}(\theta)=(X_tV(\theta)^{-1}X)^{-1}X^tV(\theta)^{-1}Y=A(\theta)^{-1}X^tV(\theta)^{-1}Y=B(\theta)Y,$$

and

$$B(\theta)^tA(\theta)B(\theta)=V(\theta)^{-1}XA(\theta)^{-1}A(\theta)B(\theta)=V(\theta)^{-1}XB(\theta).$$

Therefore it follows from (3.11)

$$\int L\left(\beta,\theta\right)d\beta=|V(\theta)|^{\frac{1}{2}}(2\pi)^{-n/2}\exp\{-\frac{1}{2}\times(Y-X\tilde{\beta}(\theta))^tV(\theta)^{-1}(Y-X\tilde{\beta}(\theta))\}\frac{(2\pi)^{p/2}}{|A(\theta)^{-1}|^{-1/2}}.$$

Then

$$\begin{aligned}l_R\left(\theta\right)=\ln\left(\int L\left(\beta,\theta\right)d\beta\right) = {} & -\frac{1}{2}\ln(|V(\theta)|)-\frac{1}{2}(Y-X\tilde{\beta}(\theta))^tV(\theta)^{-1}(Y-X\tilde{\beta}(\theta))-\frac{1}{2}\ln(|A(\theta)|)+c\\ = {} & l_p\left(\theta\right)-\frac{1}{2}\ln(|A(\theta)|)+c.\end{aligned}$$

Therefore the restricted ML (REML) of θ is given by $\hat{\theta}_{REML}$ which maximizes

$$l_R\left(\theta\right)=l_p\left(\theta\right)-\frac{1}{2}\ln(|X^tV(\theta)^{-1}X|). \tag{3.12}$$

Substituting the estimator $\hat{V}_{REML}=V(\hat{\theta}_{REML})$ into the GLS formula (3.8) gives the REML estimator of β

$$\hat{\beta}_{REML}=(X^t\hat{V}_{REML}^{-1}X)^{-1}X^t\hat{V}_{REML}^{-1}Y. \tag{3.13}$$

3.4 Confidence Intervals for Fixed Effects

In general, standard errors are often used to assign approximate confidence intervals to a parameter θ of interest. A confidence interval is often more useful than just a point estimate $\hat{\theta}$. Taken together, the point estimate and the confidence interval say what is the best guess for θ, and how far in error that guess might reasonable be. In this section we consider the way to find the confidence intervals for fixed effects under the above distribution assumptions for linear mixed models (Gaussian Mixed Models).

3.4.1 Normal Confidence Intervals

Under the assumptions that $Y \sim N(X\beta, V(\theta))$, holds and $\hat{\beta}_{REML}$ is asymptotically normal with mean 0 and approximate variance-covariance matrix given by

$$\text{Var}(\hat{\beta}) = (X^t \hat{V}_{REML}^{-1} X)^{-1}$$

where $\hat{\sigma}_i^2 = (X^t \hat{V}_{REML}^{-1} X)_{ii}^{-1}$ are considered as estimates of $\text{Var}(\hat{\beta}_i)$. Therefore

$$\hat{\beta}_i \pm z_{1-\alpha/2}\sqrt{(X^t \hat{V}_{REML}^{-1} X)_{ii}^{-1}} \tag{3.14}$$

denotes an approximate $100(1-\alpha)\%$ confidence interval for β_i.

Note that (3.14) is an approximate confidence interval, though a very useful one in a wide variety of situations (Jiang, J. (2007)). We will use the bootstrap to calculate better approximate confidence intervals, where we will be concerned on the computation of bootstrap standard errors (Efron, B. and Tibshirani, R.J. (1993)).

3.5 Bootstrapping Two-Level Models

3.5.1 Bootstrap

The Bootstrap is a technique based on simulation methods to estimate the bias and the variance (and consequently the standard error) of an estimator under minimal assumptions. In this section we discuss a little bit the general idea of the bootstrap and its applications to two-level models. It is possible to compute bootstrap estimates via the parametric and nonparametric approaches. The nonparametric bootstrap is the method where estimates are computed by generating a large number of new sample from the original sample. The parametric bootstrap method generates bootstrap observations with estimated values of parameters.

A generic implementation of the nonparametric Bootstrap, assuming sample $\{z_1, z_2, \ldots, z_n\}$ is as follows:

1. Draw B bootstrap samples $\{z_{b1}^*, z_{b2}^*, \ldots, z_{bn}^*\}$, $b = 1, 2, \ldots, B$ from $\{z_1, z_2, \ldots, z_n\}$.

2. From each of the B samples, estimate the parameter θ, thereby obtaining B estimates θ_b^*.

3 Calculate $\theta_{(.)}^* = \frac{1}{B}\sum_{b=1}^{B} \theta_b^*$,
$\hat{\text{var}}(\theta^*) = \frac{1}{B-1}\sum_{b=1}^{B}(\theta_b^* - \theta_{(.)}^*)^2$, the expectation and the variance of θ^* respectively.

4. The bias $\hat{\text{Bias}}_B = \hat{\text{Bias}}(\theta^*) = \theta_{(.)}^* - \hat{\theta}$

5. The bias-corrected of θ is $\hat{\theta}_B = \hat{\theta} - \hat{\text{Bias}}_B = 2\hat{\theta} - \theta^*_{(.)}$

In the case of two-level model,

$$y_{ij} = \underbrace{x_{ij}^t\beta}_{\text{fixed}} + \underbrace{z_{ij}^t u_i}_{\text{random}} + \underbrace{e_{ij}}_{\text{random}} , \quad i = 1, \ldots, m; \quad j = 1, \ldots, n_i, \tag{3.15}$$

el nonparametric Bootstrap (known as cases bootstrap) can be obtained by resampling entire cases as follows:

1. Draw one entire Level-2 unit (Y_i, X_i, Z_i), containing n_i Level-1 cases, with replacement,
2. From this Level-2 unit, draw a bootstrap sample (Y_i^*, X_i^*, Z_i^*) of size n_i with replacement,
3. Repeat steps 1 and 2 m times,
4. Compute all parameter estimates for the two level-model
5. Repeat steps 1-4 B times and compute bais-corrected estimates and bootstrap standard errors.

The bootstrap has proved to be an approach that derives consistent estimates of bias and standard errors of model parameters. It has been recommended to consider bootstrap estimation in multilevel models since gives satisfactory results in small sample cases under minimal assumptions. The application of the bootstrap to multilevel models is not direct because it depends on the nature the problem. During a resampling process, the hierarchical data structure must be considered since the observations are based on intra-class dependency. Several adaptations are needed in order to deal properly with the intra-class dependency. Here we will consider a two-level model and discuss one bootstrap approach: parametric bootstrap. Among the assumptions the parametric bootstrap is based on that we take into account are: The explanatory variables are considered fixed, and both the model (specification) and the distribution(s) are assumed to be correct (Efron (1982), Hall (1992), Van der Leeden, Busing and Meijer (1997)).

3.5.2 Parametric Bootstrap

The bootstrap samples are generated using the parametrically estimated distribution function of the data. In the two-level model considered here, two of these distribution functions are involved. For the level-1 residuals e, we use the $N(0, \hat{R})$ distribution function, and for the level-2 residuals, contained in the vector u, we use the $N(0, \hat{G})$ distribution function.

Let $\hat{\beta}$ be the REML estimator of β. The (re)sampling procedure is now as follows (Van der Leeden, Busing and Meijer, 1997):

1. Draw vector u^* from the a multivariate normal distribution with mean zero and covariance matrix $\hat{G}$.

2. Draw vector e^* from the a multivariate normal distribution with mean zero and covariance matrix $\hat{R}$.

3. Generate the bootstrap samples Y^* from $Y^* = X\hat{\beta} + Zu^* + e^*$.

4. Compute estimates for all parameters of the two-level model.

5. Repeat steps 1-4 B times and compute bias-corrected estimates and bootstrap standard errors.

Note that in this sampling process the covariates are assumed to be fixed.

3.5.3 Bootstrap Confidence Intervals

In many cases the bootstrap is only used for bias correction and computation of standard errors. In usual applications the standard errors are used to calculate approximate confidence interval of a parameter θ of interest. Furthermore, among other important and nontrivial applications of the bootstrap includes the computation of confidence intervals. There exist different types of bootstrap but here we only discuss one type of bootstrap confidence interval for a typical parameter θ with true value θ_0 and we will only discuss two-sided intervals. The intended nominal coverage of the confidence interval will be denoted by $1-\alpha$, so the probability that the interval contains the true parameter value should be approximately $1-\alpha$ (Efron, B. and Tibshirani, R.J. (1993)).

Bootstrap Normal confidence Interval

If the assumptions of the model, including the normality assumptions, hold, then the estimators are asymptotically normally distributed with a certain variance, derived from the likelihood function and we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\ell} N(0, \psi).$$

The distribution of $\hat{\theta} - \theta_0$ can be approximated by the normal distribution with mean zero and variance $\hat{\psi}/n$, where $\hat{\psi}$ is a consistent estimator of ψ derived from the likelihood function. The usual confidence intervals for θ_0 are therefore

$$\left[\hat{\theta} + z_{\alpha/2}\hat{\text{se}}(\hat{\theta});\ \hat{\theta} + z_{1-\alpha/2}\hat{\text{se}}(\hat{\theta})\right] \tag{3.16}$$

where $\hat{se}(\hat{\theta}) = \sqrt{\frac{\hat{\psi}}{n}}$ is the estimator of the asymptotic standard deviation of $\hat{\theta}$. Even if the random terms in the models are not normally distributed, under non harsh regularity conditions, the estimators are asymptotically normally distributed. In that case, $\hat{se}_N(\hat{\theta})$ may not be a consistent estimator of the standard deviation of the estimators of the variance components, although it is still consistent for the fixed parameters. This implies replacing $\hat{se}(\hat{\theta})$ in (3.15) by a bootstrap estimator results in an approximate confidence interval, the bootstrap normal confidence interval, which is better than the normal confidence interval.

$$\left[\hat{\theta} - z_{\alpha/2}\hat{\text{se}}_B(\hat{\theta}),\ \hat{\theta} + z_{1-\alpha/2}\hat{\text{se}}_B(\hat{\theta})\right] \tag{3.17}$$

in which $\hat{se}_B(\hat{\theta})$ is the bootstrap estimator of the standard deviation of $\hat{\theta}$. Alternatively, one might use

$$\left[\hat{\theta}_B - z_{\alpha/2}\hat{\text{se}}_B(\hat{\theta}),\ \hat{\theta}_B + z_{1-\alpha/2}\hat{\text{se}}_B(\hat{\theta})\right] \tag{3.18}$$

where $\hat{\theta}_B$ is the bootstrap bias-corrected estimator of θ (Efron, B. and Tibshirani, R.J. (1993)).

3.6 Prediction of Random Effects

3.6.1 Best Linear Unbiased Predictor (BLUP)

Technically speaking, the random effects u in model (3.3) are not model parameters like β and σ. However, as Pinheiro and Bates (2000) point out, in a way, they behave like parameters and since they are unobservable, there often is interest in obtaining estimates of their values. The best predictor (BP) of u, in the sense that it minimizes the mean squared prediction error, is the conditional mean $\tilde{u} = PB(u) = E(u|Y)$. The normality assumptions for model considered in section (3.2), imply that u and Y have a joint multivariate normal distribution

$$\begin{pmatrix} u \\ Y \end{pmatrix} \sim \mathbf{N}_{mq+n}\left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} G & GZ^t \\ ZG & V \end{pmatrix}\right), \tag{3.19}$$

and under the normal theory the conditional mean of $\mathbf{u}$ given Y is

$$E(u|Y) = E(u) + cov(u,Y)[var(Y)]^{-1}(Y - E(Y)) = GZ^tV^{-1}(Y - X\beta). \tag{3.20}$$

This is the best predictor of u, and being a linear function of Y it also is the best linear predictor (BLP) of u. In practice the unknown β in (3.19) is replaced with its estimator $\hat{\beta} = \hat{\beta}_{GLS}$, which is the BLUE of β yielding the best linear unbiased predictor (BLUP) of u:

$$\tilde{u} = BLUP(u) = GZ^tV^{-1}(Y - X\hat{\beta}). \tag{3.21}$$

The unbiasedness means here that both the random variable u and its predictor $\tilde{u}$ have the same expected value. Mathematically, the BLUP estimates of fixed parameter and random effects are defined as solution of Henderson's simultaneous equations which sometimes are called mixed model equations (Robinson, 1991).

3.6.2 Mixed Model Equations

Henderson (in Henderson et al. 1959) introduced a set of equations, solutions of which give simultaneously the GLS estimator of β and the BLUP of u. The equations are derived by maximizing the joint density of Y and u with respect to β and u.

Since $Y|u \sim N_n(X\beta + Zu, R)$ and $u \sim N_q(0, G)$, the joint density of Y and u is

$$\begin{aligned} f\,(Y,u) = f\,(Y|u)\,f\,(u) = &\ (2\pi)^{-n/2}|R|^{-1/2}\exp\{-\frac{1}{2}(Y - X\beta - Zu)^t R^{-1}(Y - X\beta - Zu)\} \\ &\times (2\pi)^{-q/2}|G|^{-1/2}\exp\{-\frac{1}{2}u^t G^{-1}u\} \\ = &\ \frac{\exp\{-\frac{1}{2}[(Y - X\beta - Zu)^t R^{-1}(Y - X\beta - Zu)\} + u^t G^{-1}u]}{(2\pi)^{(n+q)/2}|R|^{1/2}|G|^{1/2}}. \end{aligned}$$

To maximize the density $f(Y;u)$ calculate the partial derivatives of

$$\begin{aligned} \ln(f(Y,u)) = &-\frac{n+q}{2}\ln(2\pi) - \frac{1}{2}\ln|R| - \frac{1}{2}\ln|G| - \\ &\frac{1}{2}[(Y - X\beta - Zu)^t R^{-1}(Y - X\beta - Zu)\} + u^t G^{-1}u], \end{aligned}$$

with respect to β and u:

$$\frac{\partial \ln(f(Y,u))}{\partial \beta} = X^t R^{-1}(Y - X\beta - Zu)\frac{\partial \ln(f(Y,u))}{\partial u} = Z^t R^{-1}(Y - X\beta - Zu) - G^{-1}u.$$

Setting these to zero yields equations

$$\begin{aligned} X^t R^{-1}X\beta + X^t R^{-1}Zu = &\ X^t R^{-1}Y \\ Z^t R^{-1}X\beta + X^t R^{-1}Zu + G^{-1}u = &\ Z^t R^{-1}Y, \end{aligned}$$

which are written in matrix form as

$$\begin{pmatrix} X^t R^{-1}X & X^t R^{-1}Z \\ Z^t R^{-1}X & X^t R^{-1}Z + G^{-1} \end{pmatrix}\begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X^t R^{-1}Y \\ Z^t R^{-1}Y \end{pmatrix}. \tag{3.22}$$

These are Henderson's mixed model equations. Solving them produces the $\hat{\beta}_{GLS}$ and the $\tilde{u}$. A practical advantage of the mixed model equations is their computational convenience, because, unlike in (3.20), there is no need for inverting the $n \times n$ covariance matrix V. The inverses of $q \times q$ matrix G and $n \times n$ matrix R are needed instead, but they are often easy to compute: q is usually not that large and R is usually diagonal.

3.6.3 Empirical Best Linear Unbiased Predictor (EBLUP)

Usually the covariance matrices V, G and R are unknown. Then, in predicting u by the BLUP formula (3.20) they will be replaced with their REML or ML estimates to yield

$$\hat{u} = \text{EBLUP}(u) = \hat{G}Z^t\hat{V}^{-1}(Y - X\hat{\beta}). \tag{3.23}$$

The predictor $\hat{u}$ is called empirical best linear unbiased predictor (EBLUP) of u, the word empirical referring to the fact that the values of G and V have been obtained from the observed data. The estimator $\hat{\beta}$ is now the GLS estimator, where V is replaced with its estimate (i.e. $\hat{\beta}$ is typically $\hat{\beta}_{\text{REML}}$ or $\hat{\beta}_{\text{ML}}$), and it is sometimes called empirical BLUE of β. Both $\hat{\beta}$ and $\hat{u}$ can be obtained by solving the Henderson's mixed model equations where V, G and R are substituted by the corresponding estimates.

Considering the model at cluster level (level 2), model (3.2), the EBLUP for the conditional expectation $E(Y_i|u_i) = X_i\beta + Z_iu_i$ is given by

$$X_i\hat{\beta} + Z_i\hat{D}Z_i^t\hat{\Sigma}_i(Y_i - X_i\hat{\beta}). \tag{3.24}$$

The expression (3.23) may be rewritten in the form a weighted average combining the information from cluster (or subject) i only and information from the population (auxiliary information). The EBLUP for $X_i\beta + Z_iu_i$ may be viewed as borrowing strength across clusters to get the best prediction for individual i. That is, when the data from a region is small, resulting that the information from that region is weak. The regional estimate needs to be strengthened with supplementary global data. This is the fact that the smaller the regional data is, the more weight the global information gets in the estimation (Kubokawa (2009), McCulloch y Searl (2001)).

This connection between linear mixed models and small area estimation is discussed in the following chapter.

Chapter 4

Small Area Estimation with Linear Mixed Model

4.1 Introduction

In this chapter we illustrate how to apply the theory on linear mixed models and the related EBLUP theory toward the estimation of a small area characteristic of interest. To give an idea about the context, we start with the following discussion.

Suppose that the population U of size N is divided into M disjoint areas and that the survey estimates are available for m, $m \leq M$, of the areas.
Let

$$U = U_1 \cup U_2 \ldots \cup U_i \cup \ldots \cup U_M$$

and

$$N = N_1 + N_2 + \ldots + N_i + \ldots + N_M$$

where N_i is the size of area population U_i, $i = 1, 2, \ldots, M$. Assume then that a random sample s of size n is drawn from U and

$$s = s_1 \cup s_2 \ldots \cup s_i \cup \ldots \cup s_m$$

and

$$n = n_1 + n_2 + \ldots + n_i + \ldots + n_m.$$

Furthermore, each area population U_i divides into the sample s_i and the remainder $r_i = U_i - s_i$. It is possible that for some areas the sample size n_i is zero so that $r_i = U_i$.

To illustrate these concepts, we consider the following example from Fuller (2009) which considers the prediction of wind erosion in Iowa for the year 2002, where 44 counties reported data of wind erosion but the author included 4 counties with no observation due to the purpose of illustration. Each county is divided into segments, and the segments are the primary sampling units of the survey. The soils of Iowa have been mapped, so population values for a number of soil characteristics are available. The mean of the soil erodibility index for each county is the auxiliary data in the small area model.

The data set includes different variables such are county, totalsegments, samplesegments, erodibility, and Y. The variable county records an identification number for each county, totalsegments records the total number of segments in the county, and samplesegments records the sampled number of segments in the county. The variable erodibility records a standardized population mean of the soil erodibility index for each county, and Y records the survey sample mean (direct estimate) of a variable that is related to wind erosion for each county. The first 44 observations contain the observed data, and the last 4 observations contain the hypothetical counties for which there were no observations in the survey.

Relating this example to general description, the following specifications are given:

U : all segments in all counties under study
M : number of counties under the study
N : is the population (total) number of segments in all counties
U_i : all segments in the county i
N_i : the population (total) number of segments in the county (small area) i
m : number of counties that have been sampled
n : the sample number of segments in all counties
n_i : the sample number of segments in small area i
Y : wind erosion in all counties
$\bar{y}_i$: the estimate (direct estimate) of the mean of segments that have been affected by the wind erosion for the area i
$\bar{x}_i$: a vector of known means of auxiliary variables for the area i
$\bar{r}_{1i}$: the population mean erodibility index for county i

The model proposed by the author is:

$$\begin{aligned}
\bar{y}_i &= \theta_i + e_i \\
\theta_i &= \bar{x}_i^t\beta + \upsilon_i, \\
&\text{i.e} \\
\bar{y}_i &= \bar{x}_i^t\beta + \upsilon_i + e_i,
\end{aligned}$$

with $\bar{x}_i^t = [1, 0.1(\bar{r}_{1i} - 59)]$, $i = 1, 2, \ldots, 48$. Where υ_i is the area effect, e_i is sampling error,

$$\begin{aligned} \upsilon_i &\sim \text{iid } N(0, \sigma_\upsilon^2) \\ e_i &\sim \text{iid } N(0, \sigma_e^2), \end{aligned}$$

and υ_i and e_i are independent.

Due to the fact that the mean of y for area i is assumed to be the sum of $\bar{x}_i\beta$ and υ_i, the fixed and random parts respectively, this model is also called mixed model (see Chapter 3).

The model presented here is the area level model where the quantity of interest is the mean, but the nature of data may vary for different context and problems. The model can be defined in terms of primary sampling units or area level terms as in this example.

This chapter covers the statistical framework specific for small area estimation. We present present the corresponding linear mixed model in Section 4.2. In Section 4.3 we discuss Prediction for small areas quantities of interest and we close the chapter with an example from the literature.

4.2 Nested Error Regression Model

We consider a linear mixed model commonly called the nested-error regression model with a single random effect, u_i, for the *ith* small area, or small group:

$$y_{ij} = x_{ij}^t\beta + u_i + e_{ij}. \tag{4.1}$$

Here y_{ij} is the response of unit j in area i, x_{ij} is the corresponding vector of auxiliary variables, β is the vector of fixed parameters, u_i is the random effect of area i and e_{ij} the random individual error term. The area effects u_i are assumed independent with zero mean and variance σ_u^2. Similarly, the errors e_{ij} are independent with zero mean and variance σ_e^2. In addition, the u_i's and the e_{ij}'s are assumed mutually independent. Under these assumptions,

$$\mathrm{E}(y_{ij}) = x_{ij}^t\beta,$$

and

$$\mathrm{Var}(y_{ij}) = \sigma_u^2 + \sigma_e^2.$$

We also make the usual assumption that both u_i and e_{ij} are normally distributed. Let Y_i denote the $n_i \times 1$ vector of the n_i elements in area U_i, i.e.

$$Y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix}_{(n_i \times 1)} .$$

Now the model in the matrix form, for the population U_i of area i is

$$Y_i = X_i\beta + Z_iu_i + e_i, \tag{4.2}$$

where

$$X_i = \begin{pmatrix} x_{i1}^t \\ xt_{i2} \\ \vdots \\ x_{in_i}^t \end{pmatrix}_{(n_i \times p)}$$

$Z_i = 1_{n_i}$: the $n_i \times 1$ unit vector , u_i is scalar and

$$\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}_{(n_i \times 1)} .$$

With this notation $E(Y_i) = X_i\beta$ and $\text{Var}(Y_i) = V_i = J_{n_i}\sigma_u^2 + I_{n_i}\sigma_e^2$, where $J_{n_i} = 1_{n_i} \otimes 1_{n_i}^t$ is the square matrix of ones. Here the sign $\otimes$ is the *Kronecker product.*

Using the usual mixed model notation we can write $V_i = Z_iG_iZ_i^t + R_i$ where $G_i = \sigma_u^2$ and $R_i = \sigma_e^2 I_{n_i}$. If we further combine the area vectors Y_i into one response vector

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix}_{(n \times 1)} ,$$

we have the model in the form (3.3) of the general linear mixed model. The model is now

$Y = X\beta + Zu + e$, with

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}_{(n \times p)} ,$$

$$Z = \begin{pmatrix} 1_{n_1} & 0 & \dots & 0 \\ 0 & 1_{n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_{n_m} \end{pmatrix}_{(n \times m)} ,$$

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix}_{(m \times 1)}$$

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}_{(n \times 1)} .$$

Units (or observations) coming from different areas i and i' are not correlated. The covariance matrix V of the response vector Y is block-diagonal.

$$\begin{aligned} \mathrm{Var}(Y) = V = & \begin{pmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_m \end{pmatrix}_{n \times n} \\ = & \ ZGZ^t + R, \ \text{ where } G = \sigma_u^2 I_m \text{ and } R = \sigma_e^2 I_n. \end{aligned}$$

where $G = \sigma_u^2 I_m$ and $R = \sigma_e^2 I_n$.

4.3 BLU and EBLU Predictors

In this section we consider the information from unit (or individual) level to find the estimator (or predictor) of the variable of interest. We Assume that our interest is to find the total area and we suppose that $\mathbf{y} = (y_1, y_2, \dots, y_N)^t$ is a realization of a random variable Y. Let a model ξ characterizes the probability distribution of Y. The focus is to estimate the value of a population quantity of interest, which can be seen as function $h(\mathbf{y})$ of $\mathbf{y}$, typically a linear combination $\mathbf{c}^t\mathbf{y}$. If $\mathbf{c} = \mathbf{1}$, where $\mathbf{1}$ is an unit vector, then $h(\mathbf{y})$ is the population total, and if $\mathbf{c} = \mathbf{1}/N$, then $h(\mathbf{y})$ is the population mean.

Now suppose that the interest is to estimate the total population of the variable of interest for area i

$$Y_i = \sum_{j \in U_i} y_{ij}.$$

Following Royall (1976), let s denote a sample of size n from a finite population U and let r denote the nonsampled remainder of U so that $U = s \cup r$. Correspondingly, let $\mathbf{y}_s$ be the vector of observations in the sample and $\mathbf{y}_r$ the rest of $\mathbf{y}$. We have the following decomposition:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \mathbf{c}_s \\ \mathbf{c}_r \end{pmatrix}.$$

The population quantity to be estimated is now

$$h(\mathbf{y}) = \mathbf{c}_s^t \mathbf{y}_s + \mathbf{c}_r^t \mathbf{y}_r$$

a realization of the random variable

$$h(Y) = \mathbf{c}_s^t Y_s + \mathbf{c}_r^t Y_r.$$

Note that the first term $\mathbf{c}_s^t \mathbf{y}_s$ is observed from the sample, whereas the second term must be estimated (or predicted, in the frequentist terminology, since it is a function of random variables, not a fixed parameter). Thus, estimating $h(\mathbf{y})$, or predicting $h(Y)$, is essentially predicting the value $\mathbf{c}_r^t \mathbf{y}_r$ of the unobserved random variable $\mathbf{c}_r^t Y_r$.

Lets consider ξ to be a linear mixed model in a such way that

$$\mathrm{E}_\xi(Y) = X\beta$$

and

$$\mathrm{Var}_\xi(Y) = V.$$

In accordance with the partition $U = s \cup r$ we can arrange X and V so that

$$X = \begin{pmatrix} X_s \\ X_r \end{pmatrix}$$

and

$$V = \begin{pmatrix} V_s & V_{sr} \\ V_{rs} & V_r \end{pmatrix},$$

where X contains auxiliary variables, β is a vector of unknown parameters and V is an arbitrary positive definite covariance matrix. The vectors $\mathbf{y}$ and $\mathbf{e}$ as well as the matrices X and Z can be partitioned into sample parts $\mathbf{y}_s$, $\mathbf{e}_s$, X_s and Z_s (of n rows) and remainder parts $\mathbf{y}_r$, $\mathbf{e}_r$, X_r and Z_r (of $N-n$ rows). Then the nested error regression model, in the matrix form, becomes

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix} \beta + \begin{pmatrix} Z_s \\ Z_r \end{pmatrix} u + \begin{pmatrix} e_s \\ e_r \end{pmatrix}.$$

The corresponding partition of the covariance matrix of $\mathbf{y}$ is

$$V = \begin{pmatrix} V_s & V_{sr} \\ V_{sr} & V_r \end{pmatrix} = \begin{pmatrix} Z_s G Z_s^t & Z_s G Z_r^t \\ Z_r G Z_s^t & Z_r G Z_r^t \end{pmatrix} + \begin{pmatrix} \sigma_e^2 I_n & 0 \\ 0 & \sigma_e^2 I_{N-n} \end{pmatrix}.$$

The following theorem from Royall (1976) gives the best linear unbiased predictor (BLUP) of $c^t Y$ as well as its error variance under the general linear model, in the case of finite population and serves as a general basis of the BLUP approach to small area estimation with unit level models. In this context, the best linear unbiased predictor means an unbiased predictor which is linear in observed data (Y_s) and has a minimum error variance among all linear unbiased predictors.

The general prediction theorem.

Under the model ξ for a finite population U the best linear model-unbiased predictor of $h(Y) = c^t Y$ is

$$\mathrm{BLUP}(c^t Y) = c_s^t \mathbf{y}_s + c_r^t [X_r \hat{\beta} + V_{rs} V_s^{-1} (\mathbf{y}_s - X_s \hat{\beta})], \tag{4.3}$$

and the error variance is

$$\begin{aligned} \mathrm{Var}_\xi[\mathrm{BLUP}(c^t Y) - c^t Y] =& c_r^t (V_r - V_{rs} V_s^{-1} V_{sr}) c_r + \\ & c_r^t (X_r - V_{rs} V_s^{-1} X_s)(X_s^t V_s^{-1} X_s)^{-1} (X_r - V_{rs} V_s^{-1} X_s)^t c_r, \end{aligned}$$

where

$$\hat{\beta} = (X_s^t V_s^{-1} X_s)^{-1} X_s^t V_s^{-1} \mathbf{y}_s$$

is the general least squares (GLS) estimator of β.

Note that the GLS estimator is also the best linear unbiased estimator (BLUE) of β, that is, it has the minimum variance among linear unbiased estimators (McCulloch and Searle, 2001).

Proof:
As we are interested in predicting the part which is not observed, the information needed will come from observed variables (sample vector $\mathbf{y}_s$). Considering the decomposition of the population quantity of interest

$$h(Y) = \mathbf{c}_s^t Y_s + \mathbf{c}_r^t Y_r,$$

the random term is $\mathbf{c}_r^t Y_r$. Now suppose the predictor of this term is a linear function of the data, i.e. the predictor $\mathbf{c}_r^t Y_r$ is of the form $\mathbf{a}^t Y_s$ where $\mathbf{a}$ is a vector to be specified. In addition the estimator $\hat{h}(Y)$ is assumed to be unbiased. Now, define

$$\begin{aligned} \mathrm{E}_\xi(a^t Y_s - c_r^t Y_r)^2 &= \mathrm{Var}(a^t Y_s - c_r^t Y_r) + (\mathrm{E}_\xi(a^t Y_s - c_r^t Y_r))^2 \\ &= a^t V_s a + c_r^t V_r c_r - 2a^t V_{sr} c_r + [(a^t X_s - c_r^t X_r)\beta]^2, \end{aligned} \tag{4.4}$$

the last expression is obtained from the expression of V and the fact that $\mathrm{E}(Y) = X\beta$.

Now the BLUP$(c^t Y)$ will be found by minimizing the $\mathrm{E}_\xi(a^t Y_s - c_r^t Y_r)^2$ with respect to a under the constraint of model unbiasedness,

$$\mathrm{E}_\xi(a^t Y_s - c_r^t Y_r) = (a^t X_s - c_r^t X_r)\beta = 0,$$

for all β, which is equivalent to

$$a^t X_s - c_r^t X_r = 0.$$

Lets consider the following Lagrangian function

$$\mathcal{L}(a, \lambda) = a^t V_s a - 2a^t V_{sr} c_r + 2(a^t X_s - c_r^t X_r)\lambda,$$

where λ is the vector of Lagrange multipliers.

The first partial derivative respect a a is

$$\frac{\partial \mathcal{L}(a, \lambda)}{\partial a} = 2V_s a - 2V_{sr} + 2X_s \lambda,$$

equating to zero, yields

$$X_s \lambda = V_s a - V_{sr}, \tag{4.5}$$

and then

$$a = V_s^{-1}(V_{sr} c_r - X_s \lambda). \tag{4.6}$$

Now multiply (4.5) by $X_s^t V_s^{-1}$ yields $X_s^t V_s^{-1} X_s \lambda = X_s^t V_s^{-1} V_s a - V_{sr}$, using the unbiasedness constraint $a^t X_s = c_r^t X_r$ and then solving for λ, we get

$$\lambda = (X_s^t V_s^{-1} X_s)^{-1}(X_s^t V_s^{-1} V_{sr} - X_r^t) c_r,$$

then after substituting this expression in (4.6) and making some simplifications, we get

$$a = V_s^{-1}[V_{sr} - X_s(X_s^tV_s^{-1}X_s)^{-1}(X_s^tV_s^{-1}V_{sr} - X_r^t)]c_r. \tag{4.7}$$

Recall that $\hat{h}(Y) = c_sY_s + a^tY_s$. Then substituting a with (4.7), yields

$$\begin{aligned}
\text{BLUP}(c^tY) &= c_s^tY_s + (V_s^{-1}[V_{sr} - X_s(X_s^tV_s^{-1}X_s)^{-1}(X_s^tV_s^{-1}V_{sr} - X_r^t)]c_r)^tY_s \\
&= c_s^tY_s + c_r^t[V_{sr} - X_s(X_s^tV_s^{-1}X_s)_{-1}(X_s^tV_s^{-1}V_{sr} - X_r^t)]^tV_s^{-1}Y_s \\
&= c_s^tY_s + c_r^t[V_{sr}^tV_s^{-1}Y_s - (X_s^tV_s^{-1}V_{sr} - X_r^t)^t(X_s^tV_s^{-1}X_s)^{-1}X_s^tV_s^{-1}Y_s] \\
&= c_s^tY_s + c_r^t[V_{sr}^tV_s^{-1}Y_s - (X_s^tV_s^{-1}V_{sr} - X_r^t)^t\hat{\beta}] \\
&= c_s^tY_s + c_r^t[X_r\hat{\beta} + V_{sr}^tV_s^{-1}(Y_s - X_s\hat{\beta})],
\end{aligned}$$

where $\hat{\beta} = (X_s^tV_s^{-1}X_s)^{-1}X_s^tV_s^{-1}Y_s$.

The expression of error variance is found in the same way by inserting (4.7) into (4.4) •

According to the decomposition, the area total variable $Y_i = \mathbf{c}_i\mathbf{y}$ can be written as

$$Y_i = \sum_{j\in U_i} y_{ij} = \sum_{j\in s_i} y_{ij} + \sum_{j\in s_i} y_{ij},$$

where $\mathbf{c}_i$ is an $N \times 1$ vector of N_i ones and $N - N_i$ zeros such that the ones correspond to those $y_{ij}s$ of $\mathbf{y}$, which belong to area i. The same partition to $\mathbf{c}_i$ yields

$$\mathbf{c}_i = \begin{pmatrix} \mathbf{c}_{is} \\ \mathbf{c}_{ir} \end{pmatrix}$$

where $\mathbf{c}_{is}$ picks the units in the sample s_i from area i and $\mathbf{c}_{ir}$ picks those in the remainder r_i. Now the area total to be estimated is

$$Y_i = \mathbf{c}_{is}^t\mathbf{y}_s + \mathbf{c}_{ir}^t\mathbf{y}_r$$

and the general prediction theorem can be applied directly. This gives

$$\tilde{Y}_{i,\text{BLUP}} = \mathbf{c}_{is}^t\mathbf{y}_s + \mathbf{c}_{ir}^t[X_r\hat{\beta} + V_{rs}V_s^{-1}(\mathbf{y}_s - X_s\hat{\beta})],$$

where the GLS estimate

$$\hat{\beta} = (X_s^tV_s^{-1}X_s)^{-1}X_s^tV_s^{-1}\mathbf{y}_s$$

is calculated over the whole sample data s. The covariance matrix V is assumed known here.

Considering the covariance matrix decomposition where $V_{rs} = Z_r G Z_s^t$, we have

$$\tilde{Y}_{i,\text{BLUP}} = \mathbf{c}_{is}^t \mathbf{y}_s + \mathbf{c}_{ir}^t X_r \hat{\beta} + \mathbf{c}_{ir}^t Z_r [G Z_s^t V_s^{-1} (\mathbf{y}_s - X_s \hat{\beta})] = c_{is}^t \mathbf{y}_s + (\mathbf{l}_i^t \hat{\beta} + m_i^t \tilde{\mathbf{u}}),$$

where $\mathbf{l}_i^t = \mathbf{c}_{is}^t X_r$, $\mathbf{m}_i^t = \mathbf{c}_{ir}^t Z_r$ and $\tilde{\mathbf{u}}$ is the BLUP of $\mathbf{u}$ according to (3.20). Furthermore $\mathbf{l}_i^t \hat{\beta} + m_i^t \tilde{\mathbf{u}}$ is the BLUP of linear combination $\mathbf{l}_i^t \beta + m_i^t \mathbf{u}$ (Rao, 2003).

Note that $c_{is}^t y_s = \sum_{j \in s_i} y_{ij}$, $\quad l_i^t = c_{ir}^t X_r = \sum_{j \in r_i} x_{ij}^t$ and $m_i^t = c_{ir}^t Z_r$ is a row vector with $N_i - n_i$ in the entry i and zeros elsewhere. The BLUP is now

$$\tilde{Y}_{i,BLUP} = \sum_{j \in s_i} y_{ij} + (\sum_{j \in r_i} x_{ij}^t) \hat{\beta} + (N_i - n_i) \tilde{u}_i,$$

As mentioned in Chapiter 3, in practical situations, the components of covariance matrix V need to be estimated since they are not known. One of the approaches that are used include REML method. Replacing V by its estimator in GLS, we get REML estimator of β and the EBLUP of Y_i under REML estimation is

$$\hat{Y}_{i,EBLUP} = \sum_{j \in s_i} y_{ij} + (\sum_{j \in r_i} x_{ij}^t) \hat{\beta}_{REML} + (N_i - n_i) \hat{u}_i,$$

where $\hat{u}_i$ is the EBLUP of u_i obtained form (3.22).

4.4 Illustration Using Data from Battese et al. (1988)

In this section, we use the data provided in Battese, Harter and Fuller (1988). They considered data for 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and the data obtained from land observatory satellites (LANDSAT) during the 1978 growing season. They used an error components models for the prediction of the county crop areas.

Specifically, Battese et al. (1988) specified a linear regression model for the relationship between the reported hectares of corn and soybeans within sample segments in the Enumerative Survey and the corresponding auxiliary data in order to determine the areas under corn and soybeans. A nested error model, which defines a correlation structure among reported crop hectares within the counties, was used. From which the mean hectares of the crop per segment in a county was defined as the conditional mean of reported hectares, given the satellite determinations (auxiliary data) and the realized (random) county effect. The Battese et al. (1988) data showing the June-Enumerative survey and the Satellite Data is presented in Table 4.1. The data shows, specifically;

(i) The number of segments in each county,

(ii) The number of hectares of corn and soybeans for each sample segment (as reported in the June Enumerative Survey),

(iii) The number of pixels classified as corn and soybeans for each sample segment.

4.4.1 Fitting the Model

In this model fitting section we utilize the model proposed by Battese et al. (1988). However, a difference with respect to the approach of Battese et al. (1988), in this thesis, we employed the Restricted Maximum Likelihood Estimation method (REML) for estimating the model parameter, as reviewed in Chapter 3 while Battese et al. (1988) used a modified option of the SUPER CARP. Furthermore, we calculate the confidence intervals and the bootstrap interval for the fixed parameter.

According to Battese et al. (1988), the model construction is such that the reported crop hectares for corn (or soybeans) in sample segments within counties are expressed as a function of the satellite data for those sample segments. The unit level model is given by

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \upsilon_i + e_{ij}, \tag{4.8}$$

where i is the subscript for county ($i = 1, 2, \ldots, m = 12$), j is the subscript for a segment within a given county ($j = 1, 2, \ldots, n_i$, where n_i is the number of sample segments in the i^{th} county), y_{ij} is the number of hectares of corn (or soybeans) in the j^{th} segment of the i^{th} county, as reported in the June Enumerative Survey, x_{1ij} and x_{2ij} are the number of pixels classified as corn and soybeans, respectively, in the j^{th} segment of the i^{th} county, β_0,β_1, and β_2 are unknown parameters, υ_i is the i^{th} county effect and e_{ij} is the random error associated with the j^{th} segment within the i^{th} county. In addition the random effects, υ_i ($i = 1, 2, \ldots, m = 12$), are assumed to be *iid* $N(0, \sigma_v^2)$ random variables independent of the random errors, e_{ij} ($j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, m$) which are assumed to be i.i.d $N(0, \sigma_e^2)$ random variables.

Next, the sample mean for the hectares of corn (or soybeans) per segment in the i^{th} count is given by $\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ while the sample mean numbers of pixels of corn and soybeans are $\bar{x}_{1i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1ij}$ and $\bar{x}_{2i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{2ij}$ respectively. Thus model for this mean $\bar{y}_{i.}$, is given by

$$\bar{y}_{i.} = \beta_0 + \beta_1 \bar{x}_{1i.} + \beta_2 \bar{x}_{2i.} + \upsilon_i + \bar{e}_{i.},$$

where $\bar{e}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}$.

The population mean hectares of corn (or soybeans) per segment in the i^{th} county is defined as the conditional mean of the hectares of corn (or soybeans) per segment, given

Table 4.1: Survey and Satellite Data for corn and soybeans in 12 Iowa Counties.

	No.of segments		Reported hectares		No.of pixels in sample segments		Mean number of pixels per segment	
county	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamiliton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	196.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	5	556	88.59	102.59	220	262	325.99	177.05
			165.35	69.28	355	160		
			104	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

the realized county effect υ_i and the values of the satellite data. Under the assumption of the model (4.8), Battese et al. (1988) presented this mean by as

$$\bar{y}_{i(p)} = \beta_0 + \beta_1 \bar{x}_{1i(p)} + \beta_2 \bar{x}_{2i(p)} + \upsilon_i,$$

where $\bar{x}_{1i(p)} = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{1ij}$ and $\bar{x}_{2i(p)} = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{2ij}$ are the population mean numbers of pixels classified as corn and soybeans per segment, respectively, in the i^{th} county, and N_i is the total number of segments in the i^{th} county. The population mean pixel values $\bar{x}_{1i(p)}$ and $\bar{x}_{2i(p)}$ are known since the number of pixels of corn and soybeans are available from the satellite classifications for all segments in the i^{th} county (Battese et al., 1988).

By letting Y_i represents the column vector of the reported hectares of the given crop for the n_i samples segments in the i^{th} count, $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^t$. Furthermore, consider Y as the column vector of the reported hectares of the crop for the sample segments in the m counties, $Y = (Y_1, Y_2, \ldots, Y_m)^t$. Thus the model (4.8), expressed in matrix notation, in marginal form, is

$$Y = X\beta + \upsilon^*, \tag{4.9}$$

where the row of X that corresponds to the element y_{ij} in Y is $\mathbf{x}_{ij} = (1, x_{1ij}, x_{2ij})^t$ and $\beta = (\beta_0, \beta_1, \beta_2)^t$. The random vector υ^* in (4.9) combines random effects and random errors i.e $\upsilon^* = \upsilon + \mathbf{e}$. The covariance matrix for the random vector υ^* in (4.9) is given by

$$E(\upsilon^* \upsilon^{*t}) = V = \text{block diag}(V_1, V_2, \ldots, V_m), \tag{4.10}$$

where

$$V_i = J_i \sigma_\upsilon^2 + I_i \sigma_e^2$$

with J_i the square matrix of order n_i with every element equal to 1 and I_i the identity matrix of order n_i (Battese et al., 1988).

In matrix notation the mean crop hectares per segment in the i^{th} area are expressed as

$$\bar{y}_{i(p)} = \bar{x}_{i(p)}\beta + \upsilon_i, \tag{4.11}$$

where $\bar{x}_{i(p)} = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} = (1, \bar{x}_{1i(p)}, \bar{x}_{2i(p)})$.

4.4.2 Estimation and Prediction

The main focus is to predict the mean crop area (4.11) for the i^{th} county assuming that the county effect, υ_i are known. In the case where the random errors υ_{ij}^* are known, the best predictor of υ_i is the conditional expectation of υ_i, given the sample mean $\bar{\upsilon}_{i.}^*$, where

$\bar{\upsilon}^*_{i.} = n_i^{-1}\sum_{j=1}^{n_i}\upsilon^*_{ij}$ (Henderson, 1975). The expectation of υ_i, conditional on $\bar{\upsilon}^*_{i.}$, which is considered as the best predictor has the following expression

$$E(\upsilon_i|\bar{\upsilon}^*_{i.}) = \bar{\upsilon}^*_{i.}g_i, \tag{4.12}$$

where $g_i = m_i^{-1}\sigma^2_\upsilon$ and $m_i = \sigma^2_\upsilon + n_i^{-1}\sigma^2_e$. These results are due to the fact that υ_i and $\bar{\upsilon}^*_{i.}$ have a joint bivariate normal distribution.

Considering the variances components σ^2_υ and σ^2_e to be known, the β parameters of the model can be estimated by the generalized least squares estimator

$$\tilde{\beta} = (X^tV^{-1}X)^{-1}X^tV^{-1}Y, \tag{4.13}$$

which is equivalent to the ML estimator under the normal distribution of Y. Then after (4.12) a possible predictor for the i^{th} county effect, υ_i, is

$$\tilde{\upsilon}_i = \tilde{\upsilon}^*_{i.}g_i, \tag{4.14}$$

where $\tilde{\upsilon}^*_{i.} = n_i^{-1}\sum_{j=1}^{n_i}\tilde{\upsilon}^*_{ij}$ and $\tilde{\upsilon}^*_{ij} = y_{ij} - x_{ij}\tilde{\beta}$. The corresponding predictor $\tilde{y}_i$ for the county mean crop area per segment (4.11) is

$$\tilde{y}_i = \bar{x}_{i(p)}\tilde{\beta} + \tilde{\upsilon}_i, \tag{4.15}$$

for which, according to Harville (1985), is the best linear unbiased predictor of y_i.

According to the section (4.3), the predictor for the finite population mean crop hectares per segment in the i^{th} county, is given by

$$N_i^{-1}\left[\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i}(x_{ij}\tilde{\beta} + \tilde{\upsilon}_i)\right],$$

where the unobserved y_{ij} are replaced by the model predictions. It approaches the predictor (4.15) as the sampling rate decreases. Due to the fact that the sampling rates are small in this application, this expression is approximated by the predictor (4.15) which is the one of several predictors that have been suggested for the small area problem (Fay and Herriot, 1979).

Hence combining (4.14) and (4.15), a feasible predictor for the mean crop area (4.11) in county i is

$$\hat{y}_i = \bar{x}_{i(p)}\hat{\beta} + (\bar{y}_{i.} - \bar{x}_{i.}\hat{\beta})\hat{g}_i, \tag{4.16}$$

where $\hat{g}_i$, the approximately unbiased estimator for g_i, which was suggested by Fuller and Harter (1987) is such that,

$$
\begin{aligned}
\hat{g}_i &= 1 - \hat{h}_i, \\
\hat{h}_i &= [\hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2 \hat{w}_i]^{-1} [n_i^{-1} \hat{\sigma}_e^2 + (n_i^{-1} - c) n_i^{-1} \hat{w}_i], \\
\hat{m}_i &= \hat{m}_{..} + (n_i^{-1} - c)\hat{\sigma}_e^2, \\
\hat{w}_i &= 2 d_e^{-1} \hat{m}_i^{-1} \hat{\sigma}_e^4, \\
\hat{k}_i &= 2\hat{\sigma}_e^2 (\ddot{\sigma}_{ff} + n_i^{-1})^{-1} \left[\sum_{i=1}^{m} n_i b_i \right]^{-2} \left[\sum_{i=1}^{m} n_i^2 b_i (\ddot{\sigma}_{ff} + n_i^{-1})^2 \right], \\
\ddot{\sigma}_{ff} &= max[0, (m-5)^{-1}(m-3)\hat{\sigma}_e^{-2}\hat{m}_{..} - c], \\
b_i &= 1 - 2 n_i \bar{x}_{i.} (X^t X)_{-1} \bar{x}_{i.} + \bar{x}_{i.} (X^t X)_{-1} \left(\sum_{i=1}^{m} n_i^2 \hat{x}_{i.}^t \hat{x}_{i.} \right) (X^t X)_{-1} \hat{x}_{i.}^t, \\
d_i &= n_i^{-1} [1 - n_i \bar{x}_{i.} (X^t X)_{-1} \bar{x}_{i.}], \\
\hat{m}_{..} &= \left(\sum_{i=1}^{m} n_i b_i \right)^{-1} \left(\sum_{i=1}^{m} n_i u_i \right), \\
u_i &= \ \bar{y}_{i.} - \bar{x}_{i.} (X^t X)^{-1} X^t Y, \\
c &= \left(\sum_{i=1}^{m} n_i b_i \right)^{-1} \left(\sum_{i=1}^{m} n_i d_i \right), \\
d_e &= \sum_{i=1}^{m} (n_i - 1) - 2
\end{aligned}
$$

where $d_e = 22$ and $m = 12$ is the number of counties, in this application.

Under the assumptions that the random effects and error effects are normally distributed, it is easy to see that Y is normal distributed with mean $X\beta$ and Variance V, after (4.9) and (4.10).

Now, considering the models for corn and soybeans separately, we first estimate the variance components using restricted maximum likelihood estimation approach that we presented previously, and then after applying (3.13), we find the restricted maximum likelihood estimators for the β parameters . The estimators are given in tables 4.2 and 4.3 respectively.

The data used in this example were taken from Battese et al. (1988) and are presented in table 4.1.

Table 4.2: The estimated parameters for corn.

Parameters	**Estimates**
σ_v^2	140.02
σ_e^2	147.23
β_0	51
β_1	0.329
β_2	-0.134

Table 4.3: The estimated parameters for soybeans.

Parameters	**Estimates**
σ_v^2	247.53
σ_e^2	190.45
β_0	-16
β_1	0.028
β_2	0.494

Note that the variance components estimates $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ do not have the explicit expressions, they have been found after applying (3.12).

The confidence intervals for fixed effects are found using (3.14). To find the bootstrap intervals for corn and soybeans, we follow the resampling procedure shown in section (3.5.2) where $B = 1000$, replications of bootstrap, and then we apply (3.16). The results are presented in tables 4.4 and 4.5 in case of normal confidence intervals and tables 4.6 and 4.7 in the case of bootstrap confidence intervals for corn and soybeans respectively.

The predictor for the mean crop area (4.11) in county i is calculated using (4.16) and the results are presented in tables 4.8 and 4.9 for hectares of corn and soybeans respectively.

Table 4.4: Normal Confidence Interval for corn.

Parameters	**Estimates**	**lower bound**	**Upper bound**
β_0	$\hat{\beta}_0 = 51$	34.536	67.464
β_1	$\hat{\beta}_1 = 0.329$	0.295	0.363
β_2	$\hat{\beta}_2 = -0.134$	-0.171	-0.0968

Table 4.5: Normal Confidence Interval for soybeans.

Parameters	**Estimates**	**lower bound**	**Upper bound**
β_0	$\hat{\beta}_0 = -16$	-35.04	3.04
β_1	$\hat{\beta}_1 = 0.028$	-0.011	0.067
β_2	$\hat{\beta}_2 = 0.494$	0.4509	0.537

Table 4.6: Bootstrap Confidence Interval for corn.

Parameters	**lower bound**	**Upper bound**
β_0	36.767	65.709
β_1	0.2999	0.3575
β_2	-0.1667	-0.1024

This example shows the potential of small area estimation methods. They are designed to combine two types of information: We have (expensive) reliable information on the amount of land use of crops but only a few sample segments. On the other hand, we have (cheap) satellite auxiliary information on all the segments of the counties under study. By "borrowing strength" from auxiliary information from other segments we have been able to find the predicted mean hectares of corn (or soybeans) in those 12 Iowa counties.

Table 4.7: Bootstrap Confidence Interval for soybeans.

Parameters	**lower bound**	**Upper bound**
β_0	-33.916	2.736
β_1	-0.0104	0.065
β_2	0.4514	0.534

Table 4.8: Predicted Mean Hectares of corn in 12 Iowa Counties.

county	**sample segments**	**Predicted hectares**
Cerro Gordo	1	122.6
Hamilton	1	123.2
Worth	1	119.1
Humboldt	2	117.2
Franklin	3	129.9
Pocahontas	3	102.0
Winnebago	3	122.2
Wright	3	120.2
Webster	4	103.6
Hancock	5	127.7
Kossuth	5	122.0
Hardin	5	134.4

Table 4.9: Predicted Mean Hectares of soybean in 12 Iowa Counties.

county	**sample segments**	**Predicted hectares**
Cerro Gordo	1	83.1
Hamilton	1	91.3
Worth	1	91.2
Humboldt	2	95.1
Franklin	3	80.6
Pocahontas	3	113.4
Winnebago	3	87.4
Wright	3	104.5
Webster	4	112.5
Hancock	5	93.3
Kossuth	5	99.6
Hardin	5	79.3

Chapter 5

Generalized Linear Mixed Models (GLMMs)

5.1 Introduction

The linear mixed models that we revised in Chapter 3 are generally used in the situations where the observations are continuous and assume that the relationship between the mean of the dependent variable and the covariates can be modeled as a linear function. However, in some cases, the observations are correlated as well as discrete or categorical. The assumptions of a linear relationship and independence are questionable due to the nature of the dependent variable. In such cases the linear mixed models do not apply. There is a need of extensions of linear mixed models in order to consider these cases in which the observations are correlated and discrete or categorical at the same time. Such extensions can be covered by the generalized linear mixed models (GLMMs) (Jiang, 2007).

The generalized linear mixed model is the most frequently used random effects model in the context of discrete repeated measurements. Generalized linear mixed models are extensions of generalized linear models that allow for additional components of variability due to unobservable effects. Typically, the unobserved effects are modeled by the inclusion of random effects in the linear predictor of the generalized linear model (McCulloch and Searle, 2001).

The structure of the chapter is as follows: In section 5.2 we present the Generalized Linear Mixed Model with responses distributed according to a member of the Exponential family. Parameter estimation is discussed in section 5.3, where we present a simulation-based version of the Monte carlo EM algorithm.

5.2 Model Formulation

As in chapter 3, two considerations may be taken into account to define a generalized linear mixed model, such are: (i) conditional independence (given the random effects) and a conditional distribution of the random variable of interest and (ii) the distribution of random effects. Let y_{ij} be the j^{th} outcome measured for cluster (subject) i, $i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n_i$ and $\mathbf{y}_i$ is a vector of all measurements available for cluster i. It is assumed that, conditionally on random effects $\mathbf{u}_i$, assumed to be drawn independently from the $N(0, D)$, the outcomes Y_{ij} are independent with densities from the Exponential family of the form:

$$\begin{aligned} y_{ij}|u_i &\sim \text{indep. } f_{Y_{ij}|u_i}(y_{ij}|u_i) \\ f_{Y_{ij}|u_i}(y_{ij}|u_i) &= \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} - c(y_{ij}, \phi)\right\}. \end{aligned} \tag{5.1}$$

Noted that the relationship between the mean of random variable of interest and covariates is not a linear; this allows to find a transformation of the mean that is modeled as a linear model in both factors: fixed and random.

$$\begin{aligned} \mathrm{E}[y_{ij}|u_i] &= \mu_{ij} \\ \eta_{ij} = g(\mu_{ij}) &= \mathbf{x}_{ij}^t\beta + \mathbf{z}_{ij}^t u_i, \end{aligned} \tag{5.2}$$

where η_{ij} is the linear predictor and $g(.)$ is a known function, called $link function$ (since it combines together the conditional mean of y_{ij} and the linear form of predictors), with $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ vectors of known covariate values, with β a vector of unknown fixed regression coefficients, and ϕ a scale parameter. To that specification we have added u_i which is the random effects vector. Note that we are using μ_{ij} here to denote the conditional mean of y_{ij} given u_i, not the unconditional mean. To complete the specification we assign a distribution to the random effects:

$$u_i \sim f_{u_i}(u_i|D). \tag{5.3}$$

5.3 Parameter Estimation

The common approach in estimating model parameters is based on the use of the likelihood function. Unlike linear mixed models, the likelihood under a GLMM in the usual way does not have a closed-form expression due to the fact that such a likelihood may involve high dimensional integrals that cannot be evaluated analytically (Jiang, 2007). For such reason, likelihood-based inference in GLMMs is still challenging. To illustrate these difficulties lets consider the following example taken from Jiang (2007).

Let the random variable of interest conditional on the random effects have a Bernoulli distribution with the probability of success p_{ij}. This means that, given the random effects $u_1, u_2, \dots, u_{m_1}$ and $v_1, v_2, \dots, v_{m_2}$, binary responses y_{ij} are conditionally independent with

$$\begin{aligned} p_{ij} &= P(y_{ij} = 1|u, v) \\ \text{logit}(p_{ij}) &= \mu + u_i + v_j, \end{aligned}$$

where μ is an unknown parameter, $u = (u_i)_{1\le i \le m_1}$, and $v = (v_j)_{1 \le j \le m_2}$. In addition, random effects are independent and $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$, where the variances σ_1^2 and σ_2^2 are unknown. Hence the unknown parameters involved in the model which are needed to be estimated are $\theta = (\mu, \sigma_1^2, \sigma_2^2)^t$. We have:

$$\begin{aligned} f(y|u,v) &= \prod_{i=1}^{m_1}\prod_{j=1}^{m_2} f(y_{ij}|u_i, v_j) \\ &= \prod_{i=1}^{m_1}\prod_{j=1}^{m_2} \left\{ \frac{\exp\{y_{ij}(\mu + u_i + v_j)\}}{1 + \exp\{(\mu + u_i + v_j)\}} \right\} \\ f(u) &= \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^{m_1} \exp\left\{ \frac{-1}{2\sigma_1^2} \sum_{i=1}^{m_1} u_i^2 \right\} \\ f(v) &= \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right)^{m_2} \exp\left\{ \frac{-1}{2\sigma_2^2} \sum_{j=1}^{m_2} v_j^2 \right\}. \end{aligned}$$

The (marginal) likelihood function under this model is of the form:

$$L(\theta|y) = \int \dots \int f(y|u,v) f(u) f(v) du_1 \dots du_{m_1} dv_1 \dots dv_{m_2}.$$

After some manipulations the corresponding log-likelihood under this model has the following expression:

$$\begin{aligned} l(\theta|y) = &-\frac{m_1 + m_2}{2}\log(2\pi) - \frac{m_1}{2}\log(\sigma_1^2) - \frac{m_2}{2}\log(\sigma_2^2) + \mu y_{..} \\ &+ \log \int \dots \int \left[\prod_{i=1}^{m_1}\prod_{j=1}^{m_2} \{1 + \exp(\mu + u_i + v_j)\}^{-1} \right] \times \\ &\exp\left(\sum_{i=1}^{m_1} u_i y_{i.} + \sum_{j=1}^{m_2} u_j y_{.j} - \frac{1}{2\sigma_1^2}\sum_{i=1}^{m_1} u_i^2 - \frac{1}{2\sigma_2^2}\sum_{j=1}^{m_2} u_j^2 \right) du_1 \dots, du_{m_1} dv_1 \dots dv_{m_2}, \end{aligned}$$

where $y_{..} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} y_{ij}$, $y_{i.} = \sum_{i=1}^{m_1} y_{ij}$ and $y_{.j} = \sum_{j=1}^{m_2} y_{ij}$.

The above integral has no explicit expression and it cannot even be simplified.

This example shows the limitation of likelihood based inference in GLMMs. To try either to solve, or to avoid the computational difficulties, the alternative approach that can be thought is the use of Monte Carlo Expectation-Maximization (Monte Carlo EM) algorithm (McCulloch, 1997), which we present below. Before discussing the Monte Carlo EM approach, we give a brief overview of the EM algorithm.

The important point for the EM algorithm is to consider a complete data which is a combination of observed data and unobserved random variables (in this case we consider random effects). The EM algorithm is an iterative routine requiring two primary calculations in each iteration: Computation of a particular conditional expectation of the log-likelihood (E-step) and maximization of this expectation over the relevant parameters (M-step).

Let $\mathbf{y} = (y_1, y_2, \ldots, y_m)^t$ be the observed data vector and, conditional on the random effects, u, assume that the elements of $\mathbf{y}$ are independent and drawn from a distribution that belongs to the exponential family. Furthermore, assume a distribution for u depending on parameters that are elements of D.

Consider a second level model:

$$\begin{aligned} f_{y_i|u}(y_i|u) &= \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\tau} - c(y_i, \tau)\right\}. \\ u &\sim f_u(u|D). \end{aligned} \tag{5.4}$$

Here $\eta_i = x_i^t\beta + z_i^t u$, with x_i^t and z_i^t are the known vectors of covariates. The marginal likelihood corresponding to (5.4) is given by

$$L(\beta, \tau, D|y) = \int \prod_{i=1}^{m} f_{y_i|u}(y_i|u) f_u(u|D) du, \tag{5.5}$$

which cannot usually be evaluated in closed form and has an integral that is not easy to express in the explicit expression, with dimension equal to the number of levels of the random factors, u as we mentioned above. Such integrals may be integrated numerically. The goal is to develop algorithm to calculate fully parametric Maximum Likelihood (ML) estimates based on the likelihood (5.5).

The EM algorithm thus works on the augmented log-likelihood to obtain the ML estimates of the parameters over the likelihood (5.5). Now consider the class of models (5.4), where to set up the EM algorithm, we consider the random effects, u, to be the missing data.

The complete data, $\mathbf{w}$, is then $\mathbf{w} = (y, u)$, and the complete-data log-likelihood is given by

$$\ln L_w = \sum_i \ln f_{y_i|u}(y_i|u, \beta, \tau) + \ln f_u(u|D). \tag{5.6}$$

The EM algorithm then takes the following form:

1. Choose starting values $\beta^{(0)}$, $\tau^{(0)}$ and $D^{(0)}$. Set $r = 0$.
2. Calculate (with expectations evaluated under $\beta^{(r)}$, and $D^{(r)}$)
   (a) $\beta^{(r+1)}$ and $\tau^{(r+1)}$, which maximizes $E[\ln f_{y|u}(y|u, \beta, \tau)|y]$.
   (b) $D^{(r+1)}$ which maximizes $E[\ln f_u(u|D)|y]$.
   (c) $r = r + 1$.
3. If convergence is achieved, declare $\beta^{(r+1)}$,$\tau^{(r+1)}$ and $D^{(r+1)}$ to be the maximum likelihood estimators (MLE's) otherwise, return to step 2.

According to McCulloch (1997), in situations where the E-step is analytically troublesome, we may approximate an expression of this step by Monte Carlo simulations. Note this E-step is over the latent variable $\mathbf{u}$. Below we present the implementation of the Monte Carlo EM algorithm (Wei and Tanner, 1990) in which simulation methods are used to evaluate the intractable integral at the E-step. This method uses simulated random samples from the exact conditional distribution of the random effects vector $\mathbf{u}$ given the data $\mathbf{y}$, obtained via rejection sampling, using the marginal distribution of $\mathbf{u}$ as the candidate.

5.3.1 Monte Carlo EM Algorithm

The Monte Carlo EM (MCEM), introduced by Wei and Tanner (1990), is a modification of the EM algorithm where the expectation in the E-step is computed numerically through Monte Carlo simulations. The Monte Carlo estimate presents a tractable solution to problems where the E-step is not available in closed form (McCulloch (1997), Jiang (2007)).

Following McCulloch (1997), let $\mathbf{y} = (y_1, y_2, \ldots, y_m)^t$ denote the observed data with distribution $f(\mathbf{y}|\varphi)$ characterized by the s-vectors of parameters φ and $\varphi = (\beta, \tau, D)$. Consider $u = (u_1, u_2, \ldots, u_m)^t$ to be set of latent variables.
The EM algorithm thus works on the augmented log-likelihood $\ln f(\mathbf{y}, u|\varphi)$ to obtain the ML estimates of φ over the distribution $f(\mathbf{y}|\varphi)$ where it is assumed that

$$f(\mathbf{y}|\varphi) = \int f(\mathbf{y}, u|\varphi) du.$$

More specifically, the EM algorithm iterates between a calculation of the expected complete-data likelihood

$$Q(\varphi|\hat{\varphi}^{(r)}) = E_{\hat{\varphi}^{(r)}}(\ln f(\mathbf{y}, u|\varphi)|\mathbf{y}) \tag{5.7}$$

and the maximization of $Q(\varphi|\hat{\varphi}^{(r)})$ over φ, where the maximizing value of φ is denoted by $\hat{\varphi}^{(r+1)}$ and $\hat{\varphi}^{(r)}$ denotes the maximizer at the r^{th} iteration.
In particular,

$$E_{\hat{\varphi}^{(r)}}(\ln f(\mathbf{y}, u|\varphi)|\mathbf{y}) = \int \ln f(\mathbf{y}, u|\varphi) f(u|\mathbf{y}, \hat{\varphi}^{(r)}) du,$$

where $g(u|\mathbf{y}, \varphi)$ is the conditional distribution of the latent variables given the observed data and φ. In the cases where the E-step involves an integration that is difficult to evaluate analytically, we may estimate the quantity (5.6) from Monte Carlo simulations.

From this perspective, the conditional distribution of $u|y$ cannot be found in closed form. Consequently the expectations in $2a$ or $2b$ cannot be computed in closed form for the model (5.4). In addition, the marginal density f_y which is the integral of type (5.5) is not straightforward to calculate due to dimension of the vector u. There exist the possibility of generating random observations from the conditional distribution of $u|y$ by using a Metropolis-Hastings algorithm without specifying f_y. Consequently, it will be possible to approximate the expectations steps using Monte Carlo (McCulloch,1997).

If we obtain a sample $u_1^{(r)}, \ldots, u_N^{(r)}$ from the distribution $f(\mathbf{u}|\mathbf{y}, \hat{\varphi}^{(r)})$, this expectation may be estimated by the Monte Carlo sum

$$Q(\varphi|\hat{\varphi}^{(r)}) = \frac{1}{N} \sum_{k=1}^{N} \ln f(\mathbf{y}, u_k^{(r)}|\varphi). \tag{5.8}$$

To establish the Metropolis-Hastings algorithm, we first identify the candidate distribution, $h_u(u)$, from which we will generate new observations. Thereafter, we identify the acceptance function that gives the probability of accepting the new value. By letting u denote the previous draw from the conditional distribution of $u|y$, we then generate a new value, u_k^*, for the k^{th} component of u using the already specified candidate distribution. Therefore, the result is $u^* = (u_1, \ldots, u_k^*, u_{k+1}, \ldots, u_m)$. We accept u^* as the new value with probability α; otherwise, we retain u. In this case, α is given by

$$\alpha = \min\left\{1, \frac{f_{u|y}(u^*|y, \beta, \tau, D) h_u(u)}{f_{u|y}(u|y, \beta, \tau, D) h_u(u^*)}\right\}. \tag{5.9}$$

If $h_u = f_u$ is chosen, where f_u is a symmetric density function, the Metropolis-Hastings algorithm is then called the Metropolis algorithm and the equation (5.9) simplifies to

$$\begin{aligned}\frac{f_{u|y}(u^*|y,\beta,\tau,D)h_u(u)}{f_{u|y}(u|y,\beta,\tau,D)h_u(u^*)} &= \frac{\prod_{i=1}^m f_{y_i|u}(y_i|u,\beta,\tau)f_u(u^*|D)f_u(u^*|D)}{\prod_{i=1}^m f_{y_i|u}(y_i|u^*,\beta,\tau)f_u(u|D)f_u(u|D)} \\ &= \frac{\prod_{i=1}^m f_{y_i|u}(y_i|u,\beta,\tau)}{\prod_{i=1}^m f_{y_i|u}(y_i|u^*,\beta,\tau)}. \qquad (5.10)\end{aligned}$$

The Metropolis step is incorporated into the EM algorithm to give an MCEM algorithm as follows (McCulloch, 1997):

1. Choose starting values $\beta^{(0)}$,$\tau^{(0)}$ and $D^{(0)}$. Set $r = 0$.
2. Generate N values $u^{(1)}, u^{(2)}, \ldots, u^{(N)}$, from $f_{u|y}(u|y,\beta^{(r)}, D^{(r)})$ using the Metropolis algorithm described previously:
   (a) choose $\beta^{(r+1)}$ and $\tau^{(r+1)}$, to maximize a Monte Carlo estimate of $E[\ln f_{y|u}(y|u,\beta,\tau)|y]$; that is, $\frac{1}{N}\sum_{k=1}^N f_{y|u}(y|u^{(k)},\beta,\tau)$.
   (b) $D^{(r+1)}$ is chosen to maximize $\frac{1}{N}\sum_{k=1}^N f_u(u^{(k)}|D)$.
   (c) $r = r + 1$.
3. If convergence is achieved, $\beta^{(r+1)}$, $\tau^{(r+1)}$ and $D^{(r+1)}$ are declared to be the maximum likelihood estimators (MLE's) otherwise, return to step 2.

According to MacCulloch (1997), the main advantage of the Monte-Carlo EM algorithm is that it can be applied in situations where direct application of the EM algorithm is either very difficult or impossible.

Illustration using a Logit-Normal Model

We apply the MCEM algorithm to fit a simple logit-normal model from McCulloch (1997), which retains the Generalized Linear Mixed Model structure. Let $y = \{y_{ij} : j = 1, \ldots, n_i, i = 1, \ldots, m\}$ denote a set of binary response variables; here again one can think of y_{ij} as the j^{th} response for the i^{th} subject.

Let x_{ij} be a covariate (or vector of covariates) associated with the $(i,j) - th$ observation. Conditional on the random effects $U = u$, the response are independent Bernoulli(p_{ij}), where

$$\begin{aligned} y_{ij}|u_i &\sim Ber(p_{ij}) \\ u_i &\sim N(0,\sigma^2) \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta x_{ij} + u_i. \end{aligned}$$

Let $u_1, u_2, \ldots, u_m$ be independent and identically distributed as Normal$(0, \sigma^2)$. The marginal likelihood of y is given by

$$L(\beta, \sigma^2; y) = (\sigma^2)^{-m/2} \times \int_{\mathbb{R}^m} \exp\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}[y_{ij}(\beta x_{ij} + u_i) - \ln(1 + e^{\beta x_{ij} + u_i})] - \frac{1}{2\sigma^2}\sum_{i=1}^{m} u_i^2\right\} du. \tag{5.11}$$

The above expression was found by applying exponential and log functions to the joint likelihood.

We consider here a data set generated using the values of parameters proposed by Booth and Hobert (1999): With $\beta = 5$, $\sigma^2 = 1/2$, $n_i = 15, m = 10$ and $x_{ij} = \frac{j}{15}$ for each i, j.
A version of the complete data log-likelihood is given by

j i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	1	1	0	1	1	1	1	1	1	1
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1
6	0	0	0	1	0	1	1	1	0	1	1	1	1	1	1
7	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 5.1: Simulated data for the logit-normal model

$$l_w(\beta, \sigma, y, u) = -\frac{m}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m} u_i^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}[y_{ij}(\beta x_{ij} + u_i) - \ln(1 + e^{\beta x_{ij} + u_i})]. \tag{5.12}$$

To apply the EM algorithm in this problem we would need to compute the (conditional) expectation of l_w with respect to the density

$$h(u|y; \theta) \propto \exp\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}[y_{ij}(\beta x_{ij} + u_i) - \ln(1 + e^{\beta x_{ij} + u_i})] - \frac{1}{2\sigma^2}\sum_{i=1}^{m} u_i^2\right\}. \tag{5.13}$$

It is easy to see that the resulting integral will be intractable. Thus we consider a Monte Carlo EM algorithm as the means to simulate observations from the distribution given by (5.13). We apply a variable-at-a time Metropolis-Hastings independence sampler with Normal$(0, \sigma^2)$ proposals. The starting values of these runs were $(\beta^{(0)}, (\sigma^2)^{(0)}) = (2, 1)$. After applying the Metropolis algorithm for each E-step of EM we found the values of the parameters: $\hat{\beta} = 5.2$, $\hat{\sigma}^2 = 1.1$.

The following figure shows the convergence of the Metropolis algorithm.

In this chapter we presented the Generalized Linear Mixed Models and pointed out the potential complications that could happen due to the necessary integration with respect to the random effects (unobserved data). An alternative solution to this problem is discussed by presenting the Monte Carlo EM algorithm which rely on the simulation to approximate expectations in their E-step. The next chapter uses GLMMs in the context of Small Area Estimation.

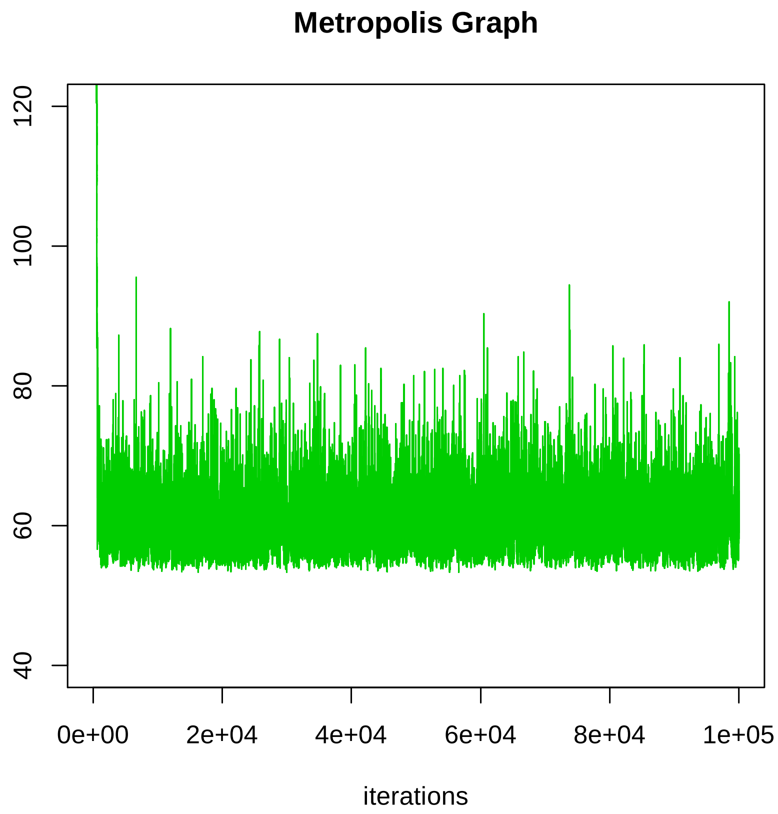

Figure 5.1: Simulated from Metropolis algorithm with Normal density proposal and conditional density $f(u|y)$ target.

Chapter 6

Small Area Estimation Under Mixed Effects Logistic Models

6.1 Introduction

In this chapter, we consider a subclass of generalized linear mixed models known as nested models with a binary response. This is basically a subgroup of the models we discussed in Chapter 5.

Chandra, Chambers and Salvati (2009) assert that the model-based methodologies allow for the construction of efficient estimators by utilizing a suitable model for borrowing strength especially with the limitation of having a small sample size. In the small area estimation context, each small area is characterized by its area-specific effect while the auxiliary information may cover all the areas of interests. Hence, the explicit linking models based on random area-specific effects that take into account between areas variation beyond the ones that are explained by auxiliary variables included in the model are used (Chandra et al.(2009), Rao (2003) and Chandra and Chambers (2009)).

When the response variables are continuous, the most common choice of inference tools to be employed include the empirical best linear unbiased predictor (EBLUP) approach under the linear mixed model (LMM) which we discussed in chapter 3. Chandra et al. (2009) considered the situation where the variable of interest is binary but the related small area estimates are needed. They used the empirical best predictor (EBP) under a generalized linear mixed model (GLMM) with logistic link function since the estimation of EBLUP is problematic when the variable of interest is binary and the sample size is small (Rao (2003), Chandra et al. (2009)).

This chapter describes a real case of discrete observations where small area techniques can be applied and how the estimation of parameters of interest can be done when the variables of interest are discretes (proportions).

6.2 Description of a Case Study

In order to illustrate the use of small area estimation context, in this section we present the model used in a study on occupational health in Canada.

6.2.1 Problem

We present an example of a real case, adopted from the article by Ghosh, Natarajan, Stroud and Carlin (1998). Their main objective was to estimate the proportion of workers who have experienced any negative impact of exposure to health hazards in the workplace for every one of the $15 \times 2 \times 2 = 60$ groups cross-classified by 15 geographic regions and the 4 demographic categories. The data presented in this article comes from a 1991 sample of all persons in 15 geographic regions of Canada. The researchers sought a response to the question "Have you experienced any negative impact of exposure to health hazards in the workplace?". In this survey, the responses were grouped into four categories: (1) yes, (2) no, (3) not exposed, and (4) not applicable or not stated. However, in this thesis we adopted only two classifications: (1) yes and (2) no, since for simplicity, we are interested in having a binary response. For each region, workers were classified by age (≤ 40 or $>$ 40) and sex (male or female). In the next subsection, we show how to find the probability of exposure to negative health hazards according to a given worker's classification using small area models.

6.2.2 Proposed Model

Following the methodology of Chandra, Chambers and Salvati (2009) let p_{ij} denote the probability that an individual j selected within the i^{th} age-sex category had experienced any negative impact of exposure to health hazards in the workplace; y_{ij} and n_i correspond to the response of workers who have been sampled and the number of individuals sampled in the workplace in age-sex group (category) i, respectively. A population model for this type of data arises when the sample counts y_{ij} are assumed to be conditionally distributed as independent binomial variables with mean $n_i p_{ij}$ with the common age-sex by category experienced any negative impact of exposure to health hazards probabilities p_{ij} following the linear logistic mixed model

$$\operatorname{logit}(p_{ij}) = \ln\left\{\frac{p_{ij}}{1-p_{ij}}\right\} = x_{ij}^t\beta + u_i = \eta_{ij}, \tag{6.1}$$

where x_{ij} is a vector of region by age-sex group level covariates and u_i is the category effect and the fixed part can be expressed as

$$x_{ij}^t\beta = \mu + \gamma_a + \gamma_s + \gamma_{as},$$

where μ is the general effect, γ_a is the main effect due to the age, γ_s is the main effect due to the sex, and γ_{as} is the interaction effect of the age and sex.

6.3 Small Area Estimation of Proportions

In this section we describe estimators for small area quantities based on generalized linear mixed models. In particular, we focus on a binary response variable with the aim of estimating the population proportions for the variable of interest in a given small area.

6.3.1 The Empirical Best Predictor for the Small Areas

When the response data are non-normal, the Generalized Linear Mixed Models (GLMMs) are the most preferred approaches for the development of indirect estimates for the small areas. The mentioned indirect estimators for small area quantities under GLMMs are usually considered as empirical best predictors (EBPs).

Now, consider U, as the finite population of size N with M non-overlapping sub-groups (or small areas) partitions U_i, each of size N_i, $i = 1, 2, \ldots, M$ such that $N = \sum_{i=1}^{M} N_i$ and $U = U_1 \cup U_2 \cup \ldots \cup U_M$. Let j and i respectively be units within small areas, where y_{ij} is the survey variable of interest (a binary variable in this case), known for sampled units, x_{ij} is the vector of auxiliary variables (including the intercept), known for the whole population. Denote s_i and r_i the sample (of size n_i) and non-sample (of size $N_i - n_i$) in small area i, respectively (Chandra et al., 2009).

The aim is to make inference about the small area quantity of interest (population proportion) in a specific area i, i.e.

$$p_i = \frac{1}{N_i}\sum_{j\in U_i} y_{ij} = \frac{1}{N_i}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} y_{ij}\right\}.$$

To make the above inference about the population proportion (p_i), we define p_{ij}, the probability that the j^{th} unit within area i, y_{ij} has the attribute of interest, u_i denotes the random area effect for the small area i, assumed to be normally distributed with mean zero and variance σ_u^2. Assuming that u_i's are independent and

$$y_{ij}|u_i \sim \text{Bin}(1, p_{ij}).$$

Therefore, a population model for the binary data is the GLMM with logistic link function, given by

$$\text{logit}\{p_{ij}\} = \ln\left\{\frac{p_{ij}}{1-p_{ij}}\right\} = \eta_{ij} = x_{ij}^t\beta + u_i, \tag{6.2}$$

with $j = 1, 2, \ldots, N_i$, $i = 1, 2, \ldots, M$, where β is the vector of regression parameters (Chandra et al., 2009).

Rao (2003) highlights that in the small area estimation, the model (6.2) can be conveniently expressed at the population level as presented next. Hence we define y_U a vector of response variable with elements y_{ij}, X_U a known design matrix with rows x_{ij}, $Z_U = \text{block diag}(1_{N_i})$ is the $N \times M$ matrix with area indicators, 1_{N_i} is a column vector of ones of size N_i, $u = (u_1, u_2, \ldots, u_M)^t$, and η_U denotes vector of linear predictors η_{ij} given by (6.1). Let $g(.)$ be a link function, such that $g(\mu) = g(E(y_U|u))$. According to the eq. (5.2) we have the linear form

$$g(\mu) = \eta_U = X_U\beta + Z_U u. \tag{6.3}$$

The expression (6.2) then defines a linear predictor for a GLMM if y_U given u are independent and belong to the Exponential family of distributions. Assume the vector of random area effects u has mean 0 and variance $G(\sigma) = \sigma^2 I_M$, where I_M is the identity matrix of order M. The link function $g(.)$ in the case of binary data is typically a logit function and the relationship between y_U (response) and η_U is expressed through a function $h(.)$, defined by $E(y_U|u) = h(\eta_U)$ (Jiang (2007), McCulloch and Searle (2001)).

Suppose we are interested in predicting a small area quantity of interest through a linear combination of the response, $\psi = a_U y_U$, where $a_U = diag(a_i^t)$ is a $M \times N$ matrix and $a_i^t = (a_{i1}, a_{i2}, \ldots, a_{iN})$ is a vector of known elements. To estimate the population proportion p_i for small area i, we consider a case where a_i^t denotes the population vector with value $\frac{1}{N_i}$ for each population unit in area i and zero elsewhere.

To express the different components of the model with respect to the sampled and non sampled parts, we decompose the vector y_U so that its first n elements correspond to the sampled units, and then partition a_U, y_U, η_U, X_U and Z_U according to the sampled and non-sampled units

$$a_U = \begin{pmatrix} a_s \\ a_r \end{pmatrix}, \; y_U = \begin{pmatrix} y_s \\ y_r \end{pmatrix}, \; \eta_U = \begin{pmatrix} \eta_s \\ \eta_r \end{pmatrix}, \; X_U = \begin{pmatrix} X_s \\ X_r \end{pmatrix}, \text{ and } Z_U = \begin{pmatrix} Z_s \\ Z_r \end{pmatrix},$$

where s denotes components defined by the n sample units and r indicates components defined by the remaining $N - n$ non-sample units. Therefore, we have $E(y_s|u) = h(\eta_s)$

and $E(y_r|u) = h(\eta_r)$. Here $h(.)$ is considered as $g^{-1}(.)$. Now the expression of quantity of interest $\psi = a_U y_U$ is represented

$$\psi = a_s y_s + a_r y_r = a_s y_s + a_r h(X_r\beta + Z_r u). \tag{6.4}$$

The first term of the right hand side in (6.4) is known from the sample, whereas the second term, which depends on the non-samples values $y_r = h(X_r\beta + Z_r u)$, is unknown and should be predicted by fitting the model (6.3) using sample data.

Let $y_s = \{y_{sij}\}$ and $y_r = \{y_{rij}\}$ denote the vector of sample values and the vector of non-samples values of the binary survey variable y respectively. The parameter of interest p_i for each small area can then be obtained by predicting each element of y_{rij}. Since we need to fit a GLMM and since it is difficult to find the explicit expression of marginal likelihood function (see chapter 5), the parameter estimation is achieved via the Monte Carlo EM technique. This gives the best linear unbiased estimate (BLUE) for β and the best linear unbiased predictor (BLUP) for u when the value of σ is assumed to be known. Using (6.3) we obtain the BLUP-type estimator of ψ. Using the estimated value $\hat{\sigma}$ of σ leads to the empirical BLUE $\hat{\beta}$ for β and the empirical BLUP $\hat{u}$ for u and thus the empirical BLUP type estimator of ψ, which is given by

$$\hat{\psi} = a_s y_s + a_r h(X_r\hat{\beta} + Z_r\hat{u}). \tag{6.5}$$

Using (6.4) the empirical best predictor (EBP) to small area i proportion p_i is then

$$\hat{p}_i = \frac{1}{N_i}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{\mu}_{ij}\right\}, \tag{6.6}$$

where

$$\hat{\mu}_{ij} = \frac{exp(\hat{\eta}_{ij})}{1 + exp(\hat{\eta}_{ij})} = \hat{p}_{ij},$$

and

$$\hat{\eta}_{ij} = x_{ij}^t\hat{\beta} + \hat{u}_i.$$

Equation (6.6) gives the population proportion estimate that we had set out to achieve (Chandra, Chambers y Salvati (2009), Saei and Chambers (2003)).

As mentioned in Chapter 5, fitting a GLMM involves evaluating a likelihood function that does not have a closed form analytical expression. It is worth noting that several approximations to this likelihood function and approximate maximum likelihood estimators have been proposed in the literature (Jiang, 2007, McCulloch and Searle, 2001). In particular, the Penalized Quasi-Likelihood (PQL) approach is a popular estimation procedure for the GLMM that is based on linear approximation to the non-normal response variable, which is then assumed to have an approximately normal distribution. This approach is reliably convergent but tends to underestimate variance components as well as fixed effect coefficients (Breslow and Clayton, 1993). Saei and Chambers (2003) described an iterative procedure to obtain the Maximum Penalized Quasi-Likelihood (MPQL) estimates of β and $\mathbf{u}$ for given σ. They assert that at convergence, the MPQL estimate of ψ is obtained by substituting the converged values of β and $\mathbf{u}$. However, in practice the variance component parameter σ is unknown and is estimated from sample data. One of the disadavantages of the MPQL approach is that, the variance component estimator is biased thus limiting its usage in the practice (Saei and Chambers, 2003). To address this shortcoming, the parameter estimates can be found using the ideas that we presented in chapter 5, where it is shown that incorporating a Metropolis-Hastings step allows construction of a MCEM algorithm for ML in GLMMs (Jiang (2007), McCulloch and Searle (2001), Saei and Chambers (2003)).

Chapter 7

Concluding Remarks and Future Research

In this work we have considered model-based approaches to finding the estimates of small area target quantities. We employed statistical models that "borrow strength" in making estimates of mean hectares of corn (or soybeans) in 12 Iowa counties where the small areas of interest had sample sizes that varied from 1 to 3. Two types of information are considered; information on the amount of land use for crops but only on a few sampled segments and satellite information on all the segments of the counties under study, which constituted the auxiliary information.

The linear mixed model was used to find the estimates for the regression coefficients and the variance components using the Maximum Likelihood Estimation (MLE) and the Restricted Maximum Likelihood Estimation (RMLE). However, the MLE was found to be biased and so to calculate the variance components estimators, we recommend the use of REML method instead. Furthermore, we calculated the confidence intervals for fixed effects, both normal confidence intervals and bootstrap normal confidence intervals.

Under the GLMMs, due to the complexity of the (marginal) likelihood function, the use of ML and REML estimates was limited by lack of explicit expressions and so we adopted the Monte Carlo EM algorithm for maximum likelihood fitting of generalized linear mixed models which uses simulation to construct Monte Carlo approximations at the E-step. For each E-step, the Metropolis algorithm was used. The resulting parameter estimates were found to be consistent with the model parameters used during the simulation. It is also worth to note that the Metropolis algorithm converged pretty fast.

This work confined the attention to the the framework of mixed models with univariate normal random area-specific effects. However, in real life, this assumption may not always

be justified. It would be a rewarding topic for future work to assess the performance of those models under multivariate normal random area effects.
Other recommendations on future research include:

- To consider the application of generalized linear mixed model in the context of small area using real data.
- To set up and study methods of small area estimation under LMMs or GLMMs with temporal and spatial effects.

Bibliography

[1] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28-36.

[2] Box, G.E.P and Jenkins, G.M (1970). *Time Series Analysis Forecasting and Control*, San Francisco: Holden-Day, Inc.

[3] Brackstone, G.J. (1987). Small area data: Policy issues and technical challenges. In *Small Area Statistics* (R. Platek, Rao J. N.K , C. E. Sarndal and M. P. Singh eds.) 3-20. Wiley, New York.

[4] Brady, T.W., Welch, K.B. and Andrzej, T.G. (2007). *Linear Mixed Models. A Practical Guide Using Statistical Software.* Chapman and Hall, Taylor and Francis Group, New York.

[5] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, pp. 9-25.

[6] Chandra, H., Chambers, R. and Salvati, N. (2009). Small area estimation of proportions in business surveys, *Centre for Statistical and Survey Methodology*, The University of Wollongong, Working Paper.

[7] Cressie, N.A. (1991). *Geostatistical analysis of spatial data. Spatial Statistics and Digital Image Analysis.* Washington, D.C.: National Academy Press.

[8] Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation, *Annals of Statistics*, 19, 1748-1770.

[9] Datta, G.S., Kubokawa, T. and Rao, J.N.K. (2002). *Estimation of MSE in Small Area Estimation*, Technical Report, Department of Statistics, University of Georgia, Athens.

[10] Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, New York.

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

[12] Efron, B. (1982). *The Jackknife, the Bootstrap and Resampling Plans*, Philadeliphia: SIAM.

[13] Erciulescu, A.L. and Fuller, W.A. (2013). Small area prediction of the mean of a binomial random variable, *JSM Proceedings*. Survey Research Methods Section. Alexandria, VA: American Statistical Association, 855-863.

[14] Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 269-277.

[15] Fuller,W.A. (2009). *Sampling Statistics*, John Wiley and Sons, Hoboken, New Jersey.

[16] Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation, *Journal of American Statistical Association*, 93, 273-282.

[17] Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates, *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 33-36.

[18] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer.

[19] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.

[20] Harville, D.A. (1985). Decomposition of prediction error, *Journal of the American Statistical Association*, 80, 132-138.

[21] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*,75, 423-447.

[22] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*, Springer, New York.

[23] Kubokawa, T. (2009). A review of linear mixed models and small area estimation, *University of Tokyo*, CIRJE Discussion Paper. http// www.e.u.tokyo.ac.

[24] Larsen, M.D. (2003). Estimation of small-area proportions using covariates and survey data, *Journal of Statistical Planning and Inference*, 112, 89-98.

[25] Longford, N.T. (2005). *Missing Data and Small Area Estimation.* Springer, New York.

[26] Malec, D. (2005). Small area estimation from the american community survey using a hierarchical logistic model of persons and housing units, *Journal of Official Statistics*, 21, 3, 411-432.

[27] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association*, 92,162-170.

[28] McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*, New York: Wiley.

[29] Montanari, G.E., Ranalli, G.M. and Vicarelli, C. (2010). A comparison of small area estimators of counts aligned with direct higher level estimates, *Scientific Meeting of the Italian Statistical Society* http://homes.stat.unipd.it/mgri/SIS2010/Program/contributedpaper/678-1393-1-DR.pdf.

[30] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer.

[31] Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Interscience .

[32] Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects, *Statistical Science*, 6, 15-32.

[33] Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling, *Journal of the American Statistical Association*, Vol. 71, No. 355, pp. 657-664.

[34] Saei, A. and Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. *Methodology Working Paper- M03/15,*, University of Southampton, United Kingdom.

[35] Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*, John Wiley and Sons, New York.

[36] Van der Leeden, R., Busing, F.M.T.A. and Meijer, E. (1997). Bootstrap methods for two-level models. *Technical Report* PRM 97-04, Leiden University, Department of Psychology, Leiden.

[37] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, Berlin.

[38] Vizcaino, L.E., Cortina-Lombardia, M.J. and Morales-Gonzalez, D. (2011). Multinomial-based small area estimation of labour force indicators in Galicia, *X Congreso Galego de Estatística e Investigación de Operacións Pontevedra.*