# Some Diagnostic Techniques for Small Area Estimation: With Applications to Poverty Mapping

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

**Doctor of Philosophy**

**in**

**Statistics**

at Massey University, Palmerston North, New Zealand.


**Alison Livingston**

2019

## Abstract

Small area estimation (SAE) techniques borrow strength via auxiliary variables to provide reliable estimates at finer geographical levels. An important application is poverty mapping, whereby aid organisations distribute millions of dollars every year based on small area estimates of poverty measures. Therefore diagnostics become an important tool to ensure estimates are reliable and funding is distributed to the most impoverished communities.

Small area models can be large and complex, however even the most complex models can be of little use if they do not have predictive power at the small area level. This motivated a variable importance measure for SAE that considers each auxiliary variable's ability to explain the variation in the dependent variable, as well as its ability to distinguish between the relative levels in the small areas. A core question addressed is how candidate survey-based models might be simplified without losing accuracy or introducing bias in the small area estimates.

When a small area estimate appears to be biased or unusual, it is important to investigate and if necessary remedy the situation. A diagnostic is proposed that quantifies the relative effect of each variable, allowing identification of any variables within an area that have a larger than expected influence on the small area estimate for that area. This highlights possible errors which need to be checked and if necessary corrected.

Additionally in SAE, it is essential that the estimates are at an acceptable level of precision in order to be useful. A measure is proposed that takes the ratio of the variability in the small areas to the uncertainty of the small area estimates. This measure is then used to assist in determining the minimum level of precision needed in order to maintain meaningful estimates.

The diagnostics developed cover a wide range of small area estimation methods, consisting of those based on survey data only and those which combine survey and census data. By way of illustration, the proposed methods are applied to SAE for poverty measures in Cambodia and Nepal.

# Acknowledgements

A PhD is not done in isolation. There are numerous people who have provided knowledge, academic support, financial support, and emotional support along the way and without these people this PhD would not have been possible.

I would like to express my deepest appreciation to my supervisors Professor Stephen Haslett and Professor Geoff Jones. Thank you for the framework for this research and providing me with opportunities to travel and be involved in interesting projects. I am forever grateful for the time and knowledge you have dedicated to help me along, as well as your patience, kindness and understanding. I am blessed to have two supervisors with a broad sense of both knowledge and kindness. This PhD thesis would have not been possible without your extensive support and encouragement.

I would like to acknowledge and thank the World Food Programme (WFP) – Nepal and Central Bureau of Statistics, Nepal, along with the WFP – Cambodia and National Institute of Statistics, Cambodia for the availability of the data sets.

I am grateful to my fellow postgraduate friends. Thank you for making days and occasionally the long nights far more bearable and enjoyable.

Thank you to Massey University and the Institute of Fundamental Sciences for the provision of scholarships and providing travel grants.

Thank you to all the academic and non-academic staff in the Institute of Fundamental Sciences who have provided advice, support, encouragement and help over the years. It has been very much appreciated.

My sincere thanks to my family and friends who have encouraged and supported me along the way and made life outside of study enjoyable.

My sincere thanks and gratitude goes to my number one supporter and biggest cheerleader, Sam. Thank you for your constant support, for wiping away the tears and having more faith in me than I had in myself.

The biggest thank you goes to God, the giver of all knowledge and wisdom.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Poverty eradication and alleviation are global objectives that have had increasing attention over the past three decades. In September 2000 world leaders met at the United Nation head quarters to set global goals with an aim to improve the standard of living in the world. There were eight main goals, with each of these having a time bound target of achieving the goal by 2015. These became known as the Millennium Development Goals (MDG). In particular, the first Millennium Development Goal was to eradicate extreme poverty and hunger; with the measurable objective being to halve the people living on less than the equivalent of US$1.25 (purchasing power parity) between 1990-2015. In order to efficiently achieve this target, aid organisations require reliable regional estimates of the poverty level in order to target support to households with the greatest level of deprivation. It would not be effective to distribute resources and aid to every single person in a country. Rather it would be more effective to target aid to the most vulnerable people to help them gain access to the resources that would benefit them the most. Other than doing a census, which is very expensive, it is difficult to gather information on all members of a population. This is where small area estimation can become useful as it can estimate poverty rates of communities of people without having to directly measure each person's

level of deprivation. This can be used to target the funding to small areas with the greatest deprivation.

The demand for small area estimation (SAE) has increased substantially in recent years. Organizations from both the private and public sectors require information at finer levels of aggregation in order to make effective decisions and implement actions specific to small areas. Along with SAE for poverty estimation, there is a much wider range of applications, including but not restricted to health, nutrition, economics and agriculture. The need for SAE arises as the sample size collected in each small area is seldom large enough to produce reliable results using direct estimation, where direct estimation uses only values of the variables of interest from the sample units in that area (Rao and Molina, 2015). The small sample size in each area produces standard errors that are unacceptably large, hence the estimates are not meaningful. In this case, supplementary data is required in order to "borrow strength" from other small areas or data sources. This in turn generates estimates with smaller levels of uncertainty surrounding them and therefore are more useful.

While such small area estimation techniques have been developed and used extensively over the last twenty years, diagnostics are not so well researched. This is especially so for the types of linear and non-linear mixed models used in small area estimation of poverty. In these applications, a statistical model is first fitted to sample survey data and then used to produce predictions assisted by census data, which are then aggregated to small area level. There are several model based methods that can be used to generate these small area estimates, including the Elbers, Lanjouw and Lanjouw (ELL) method, the Molina and Rao Empirical Bayes, Hierarchical Bayes and the M-quantile method. Such estimates have been used for the allocation of billions of dollars worth of aid; for this reason it is imperative that small area estimates are reliable so the funding is going to people with the greatest deprivation. In poverty mapping applications in SAE there maybe more than one goal, for example the need is not just for unbi-

ased estimates but also for precise estimates as well as reliable ranking of the estimates. This is related to the 'triple goal estimation' of Shen and Louis (1998), where the triple goals are: good estimates, good ranks and a good histogram (i.e good estimates of the distributional structure). This gives the motivation to investigate diagnostics of SAE.

Commonly in SAE, diagnostic techniques are applied to the 'training' data that the regression model is fitted to, with a large amount of the previous research focused on reducing the mean square error. However, the diagnostics presented in this thesis take into account the effect not only on the 'training data' but also on the small area estimates.

## 1.1 My Thesis Contribution

This thesis makes three main contributions to small area estimation diagnostics, with a particular focus on applications to poverty mapping when survey data is supplemented with unit level census data. However, these methods can be adapted to other small area applications when unit level census data is not available. The diagnostics presented focus on the final small area estimates, rather than the fit of the regression parameters, which traditional diagnostics tend to focus on.

The first contribution proposes a variable importance measure for small area estimation. The proposed metric assists in determining the variables that may not be useful in contributing to the final small area estimates. After determining the importance of the variables at the small area level, variables are sequentially removed to measure the impact on the small area estimates. This helps to determine firstly, how important a particular variable is in the presence of other variables included in the model, and secondly the magnitude to which a model can be simplified without significantly altering the final area estimates.

The second contribution introduces a specific diagnostic to measure the influence a variable has on a small area estimate. The influence measure combines the regression coefficient with the difference between a small area mean and a global mean, or a more localised mean for a particular variable. Although diagnostics are very common in linear models, they are largely neglected in small area estimation, or are only applied to the 'training' data to determine which values are having a large influence on model parameters. Rather this measure uses the supplementary data such as the census data to help determine which variable(s) are driving anomalous small area estimates.

The third contribution assesses the minimum level of precision needed in order to generate reliable small area estimates. This is done by defining a new measure $\kappa$, that takes into account the ratio of the between small area variation to the average standard error of the small areas. The reliability of the estimates are assessed using the Pearson and Spearman correlation, where the estimates are compared to a set of 'true' small area statistics. It is essential in SAE that estimates are reliable and precise enough for their purpose. Generally in model fitting exercises the fit of the model is tested via the fit of $R^2$. If a model is not explaining a large amount of variation then it may not provide precise enough estimates to be useful. However when generating the final level small area estimates, the model fit is not the only source of variability to consider. Not only is there the uncertainty in the fit of the model, but also there is variability within the small areas and variability between the small areas. In order to generate meaningful small area statistics the ranking and the values of the estimates should reflect the value and the ranking of the true small area statistics. This would correspond to the uncertainty in the estimates being small compared to the differences between them.

## 1.2 Thesis Outline

This thesis is organised into eight chapters; with three of these being contributions of the thesis. Chapter two and three review relevant literature, giving an overview of the research done so far in regards to small area estimation in general. In particular chapter two summarizes the literature related to small area estimation. In SAE, the data used to form the regression model is typically collected from a sample survey. This chapter reviews statistical techniques to adjust for the consequences of using complex sample data as well as variance estimation techniques. Additionally, it explores some popular SAE methods used when the aim is to model a linear function of the mean. Finally, it reviews current diagnostics used for small area models.

Chapter three focuses on poverty and particularly SAE poverty applications. It reviews the measures of poverty estimation with a particular focus on the Foster, Greer and Thorbecke (FGT) method to measure the poverty incidence, gap and severity. Additionally, it reviews and compares SAE techniques that are specifically designed for poverty. These include the World Bank method, otherwise known as the ELL method, the Empirical best predictor, also known as the Molina and Rao Empirical Bayes method (EB_MR), the Hierarchical Bayes and the M-quantile method.

Chapter four introduces the data sets that are used for the applications in the remainder of the thesis. These include the 2008 Cambodian unit level census data and the Cambodian Socio-Economic survey collected in 2009. These two data sets were used to generate the small area poverty estimates in Cambodia. The model and data from this poverty mapping exercise are used in Chapters five and seven to illustrate the methodologies introduced. Data from Nepal is then described, including the unit level census data collected in 2010 and the Nepalese health survey collected in 2011; this data is used in Chapter six.

Chapter five outlines the first contribution to the thesis. In general small area models can

be very large, for example have 30+ variables. Due to the often limited availability of variables, many of those included are categorical predictor variables. This is especially common in cases when survey and unit level census data are available such as when the ELL model is applied. The variables included in the model are deemed as significant at the model fitting stage, where the model is fitted using the training data. However, it has not been tested if these variables are important at the small area level. Therefore the variable importance measure will help determine which variables are important in producing meaningful small area estimates. This chapter applies the proposed methodology and the theory of this variable importance measure to a poverty mapping application in Cambodia.

Chapter six outlines the thesis's second main contribution. The chapter investigates how to determine if any observations or variables are having a large influence on a small area estimate. This method can be applied if a small area is seen to be particularly unusual or to see if there is any particular variable(s) that are driving this distinctive difference. Alternatively, it can be used as a diagnostic to examine if any particular variables in small areas are unusual. This chapter was motivated by a poverty mapping exercise in Nepal in which a particular small area was seen as having an exceptionally high rate of children being underweight, and a diagnostic needed to be developed to observe what was driving this rate.

In small area estimation, estimates need to be at a reasonable level of accuracy in order to be useful. Chapter seven outlines the third contribution where it proposes a measure to take into account the uncertainty within a small area as well as the variation between the small areas. This will help to determine the level of accuracy small area estimates need to attain in order to be useful. A simulation study is used to measure the level of accuracy needed in general to be able to produce reliable estimates. In small area estimation, models are seldom linear with a normal distribution, so while measuring the accuracy needed for a normal distribution, a skewed distribution is also examined as well a non-linear transformation. The theory is then applied to

the Cambodian poverty data to examine if the results hold for a real data application.

A summary of the work and the conclusions are given in chapter eight as well as the future work to be done.

# Chapter 2

# Small Area Estimation

## 2.1 Introduction

Small area estimation can be defined as a statistical technique that estimates parameters for sub populations (Rao, 2003). The need for SAE has arisen from the demand for reliable estimates at finer levels of aggregation for a target population. It is not always feasible to perform a full enumeration census as it is both expensive and time consuming, therefore sample surveys are often conducted to answer a question and gain reliable information for a population of interest. Although surveys are cost effective they are not able to provide reliable estimates at finer levels of the population, due to the sample size being insufficient or non-existent. This is where SAE becomes an important tool, as it allows reliable estimation to occur at small area level where direct estimation is not possible, for example estimating the unemployment rate at territorial local authority level area in New Zealand (Haslett et al., 2008). Generally a sample survey is collected and regression parameters are estimated from this, the model is then applied to a large survey or census. It is assumed that the model fitted to the sample survey holds for the larger data set as well. Area level effects or sub-population effects are added to account for

differences between the areas. Furthermore more contextual variables are often included to explain variation between the small areas.

Small areas are often thought of as geographic domains such as regions, districts or municipalities, however they can also be cohorts of people, such as different socio-economic groups or individuals of differing age ranges. The SAE method applied often depends on the structure, quantity and availability of data. Different techniques have been developed, some model only survey data, while others model survey data and combine this with census data or another large data source. Implicit models such as indirect domain estimation using synthetic estimation and composite estimation can be used to improve estimates by borrowing strength. Alternatively, explicit models using auxiliary information can be used to account for the between area variation. Or some combination of the two can be used. Small area estimation can be separated into design based direct and indirect estimation and model based direct and indirect estimation. In general, model based indirect estimation offers several advantages; that are outlined in Rao and Molina (2015). There has been particular focus on empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB). These methods are useful for modelling linear functions of the mean and are extensively outlined in Rao (2003), Rao and Molina (2015) and Ghosh and Rao (1994) as well as there being overviews in a number of other SAE literature.

The research on SAE is extensive, with the main reference being Rao (2003) and its update Rao and Molina (2015), which give a detailed overview of both design and model based methods. Rao (2003) is itself an updated and more detailed version of Ghosh and Rao (1994). Pfeffermann (2002, 2013) also provides reviews of the developments in SAE. A consequence of this rapid and extensive development is that not a single resource contains information on all the current methods. Despite the extensive coverage in Rao and Molina (2015) there are papers providing more in-depth explanations of specific methods or aspects of a method that

are not covered in detail elsewhere. An example is the ELL method of Elbers et al. (2003) or the Empirical Best Prediction method in Molina and Rao (2010).

Depending on the availability of data, small area estimation can either use area level models or unit level models; these are explained in Section 2.2.1. For example, in poverty mapping, unit level models are used since information is available at household or person level. Whereas if information is not available at household level but rather area level, then area level models are used. In cases where administrative data is only available at higher levels of aggregation, the area level model is used. These models are further defined in Section 2.2.1. Furthermore the structure of the data will influence the initial regression model fitted to the data. For example, a linear model may be fitted, but usually it is some form of linear mixed model or generalised linear mixed model, as this allows for area level variation; as outlined in Section 2.2. From here the model fitting techniques need to be incorporated to ensure the model parameters are not biased. When a mixed model is used the variance components associated with the random effects need to be estimated; as outlined in Section 2.3.1. Furthermore a complex sample tends to be used to collect the data that the model is to be fitted to; as outlined in Section 2.3.2. In addition, when working with complex survey data a simple formula is not always available to estimate the variance; these estimation techniques are outlined in Section 2.3.3. After the model is fitted, there are several popular methods that can be used to produce the small area estimates, including the Empirical Best Linear Unbiased Predictor, Empirical Bayes, Hierarchical Bayes and M-quantiles; a brief explanation of these will be outlined in Section 2.4. Finally, once the model is fitted it is important to perform model diagnostics to ensure model assumptions hold and the estimates are valid; Section 2.5 outlines various current diagnostic methods.

## 2.2 Models in Common Use for SAE

Depending on the availability of the auxiliary information, small area models can either be at unit level or area level. Furthermore small areas models are often a specified version of linear mixed models (LMM) or generalised linear mixed models (GLMM), where the random effects are used to account for the specific differences between the areas that are not included explicitly via auxiliary or contextual variables. The following provide a framework for the model based methods used for small area estimation.

### 2.2.1 Area and Unit Level Models

Ghosh and Rao (1994) classify small area models into two main categories, namely area level and unit level models. Area level models are used when the auxiliary variables are only available at area level. The model gives estimates for each small area $i$, where $i = 1,...,I$, using the area level covariate information ($\mathbf{x}_i$). The original area level model proposed for small area estimation was introduced by Fay and Herriot (1979). This is defined as

$$\tilde{y}_i = \theta_i + e_i; \quad \theta_i = x_i'\beta + u_i \tag{2.1}$$

where $\tilde{y}_i$ is the direct survey estimate based only on the data in area $i$, $\theta_i$ is the true value, $e_i$ is the survey error, which has known structural properties based on the sampling design, and $u_i$ is the model error which is assumed to have a mean of zero and variance $\sigma_u^2$ independent of $e_i$. The small area estimate for area $i$ is a weighted average of the direct estimate $\tilde{y}_i$ and the model-based prediction $\hat{\theta}_i = x_i'\hat{\beta}$.

The second class of small area models can be used when unit level auxiliary data are available. The initial model proposed in small area estimation was by Battese et al. (1988).

This nested model assumes that the response for an individual unit $k$ in small area $i$ is given by

$$y_{ik} = \mathbf{x}'_{ik}\beta + u_i + e_{ik}; \qquad i = 1,...,I; \quad k = 1,...,n_i \tag{2.2}$$

where $\mathbf{x}_{ik} = (x_{ik1},...,x_{ikp},...,x_{ikP})'$ is a vector of auxiliary information at unit level, $n_i$ is the sample size in the $i^{th}$ small area and the individual level error is $e_{ik} = h_{ik}\tilde{e}_{ik}$; where $h_{ik}$ are known constants dependent on the survey design. It is often assumed that the survey errors $\tilde{e}_{ik}$ and the random effects $u_i$ are normally distributed and independent with an expected mean of zero.

In poverty mapping applications this model can be adapted to include cluster level random effects $v_j$, however there are rarely both small area level and cluster level random effects included in the model, instead either the small area level effects $u_i$ or the cluster effects $v_j$ are used and the error redefined as $e_{jk}$. Additionally in SAE applications of child stunting, undernutrition and wasting the model can be extended to a two-stage nested model: as well as fitting a small area level model effect, or a cluster level effect, a household level effect is included, which is common to every child in the household as well as the unit record child effect.

When the primary sample units (psu) are at the area level of interest then $u_i$ should capture the unexplained area-level variation, however an additional psu level error term may be required if the small areas contain several psus. The auxiliary data is not restricted to unit level data from the sample but can also include aggregate population means, such as cluster or primary sampling unit means within a small area $i$. Non-sampled means that can be obtained from the census data can be included; these are often called contextual variables. Haslett (2016) outlines that including variables that are at higher levels of aggregation compared to just unit level can markedly reduce the variation of random effects, thereby reducing the bias if there are terms omitted from the model.

### 2.2.2 Linear Mixed Model

Linear mixed models are commonly used in SAE, as they account for between area variation as well as the within area variation. The LMM can be used provided the dependent variable is continuous and errors are normally distributed. An example of an LMM is the nested unit level model developed by Battese et al. (1988) in (2.2).

The Linear Mixed Model (LMM) can be defined as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{e} \tag{2.3}$$

where the response variable $\boldsymbol{y}$ is a vector that is linearly related to a matrix of covariates $\mathbf{X}$ through regression covariates $\beta$. The second term of the model consists of the random component of the model where $\mathbf{Z}$ is a matrix of known covariates and $\mathbf{v}$ are the unobservable random effects. The last component, $\mathbf{e}$, is a vector of model errors. We assume that the unobservable components such as the random effects $\mathbf{v}$ and the error term $\mathbf{e}$ are independent, with an expected mean of zero and they are assumed to have finite variances. Furthermore, if normality is assumed then the linear mixed model is called a Gaussian linear mixed model (Jiang, 2010). The covariance matrices for $\boldsymbol{v}$ and $\boldsymbol{e}$ are commonly defined as $\mathbf{G} = V(\mathbf{v})$ and $\mathbf{R} = V(\mathbf{e})$. The variance of $\mathbf{y}$ can be defined as $V(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, where $\mathbf{Z}\mathbf{G}\mathbf{Z}'$ explains the between small area covariance and $\mathbf{R}$ is the covariance within the small areas.

Introducing the random effects can account for different correlation structures between the small areas or clusters. The simplest correlation structure is in the form of the block covariance structure, where all the units in the cluster have the same random effects, as explained in Rao (2003). The use of the LMM in small area estimation is equivalent to predicting the unobserved area specific random components for the superpopulation (Saei and Chambers, 2003); here the actual finite population is assumed to be a random realization from a conceptual "superpopu-

lation" (Malec, 2008) with an infinite number of units within an infinite number of clusters. A superpopulation refers to the situation when the sample is taken from a finite population.

In LMM there can be issues with model selection, such as difficulty in identifying the degrees of freedom due to lack of independence between observations. Müller et al. (2013), Jiang et al. (2008) and Datta et al. (2011) outline methods for model selection in linear mixed models. Furthermore Xu (2003) investigates methods to measure the explained variance in LMM, through adapting the $R^2$. Their work is not directly relevant in this thesis, as the focus is on diagnostics for the final small area estimates, not on the model fitted to the training data.

### 2.2.3 Generalised Linear Model

In small area estimation the variable of interest is not always linearly related to the matrix of covariates, hence the assumptions of the linear model or LMM do not hold. Furthermore the response may not be continuous but instead discrete or binary, so that inference which is based on the linear model is not valid. In this situation we can apply the generalised linear model (GLM) or generalised linear mixed model (GLMM). The GLM can only be used when there are only fixed effects, if there are random effects the model assumptions will no longer hold as the error terms will not be independent. In this case the GLM can be extended to a GLMM, which accounts for both fixed and random effects. This model was first considered for SAE by MacGibbon and Tomberlin (1989) as cited by Pfeffermann (2013) to use in the application of small area estimation, and is similar to (2.3), however instead of modelling the response $y$ we now model a function of the response $g(.)$. Noble et al. (2002) as well as Saei and Chambers (2003) have explored the use of GLM for small area estimation.

GLMs were originally formulated by Nelder and Wedderburn (1972). GLMs consists of three specific components:

i) the probability distribution;

ii) the link function;

iii) the linear predictor.

The form of the generalised linear model is $g(\mu_i) = x_i\beta$ where $\mu \equiv E(Y)$. Here $g(.)$ is a monotonic and differentiable link function, which could be one of a variety of functions such as the log, identity or power function. The exponential family of distributions is defined by the class of density functions given in (2.4)

$$f_i(y_i|\lambda,\phi) = \exp\left\{\frac{y_i\lambda_i - b(\lambda_i)}{a_i(\phi)} + c_i(y_i\phi)\right\}. \tag{2.4}$$

In this case $\lambda$ is associated with the mean of the distribution i.e $\mu_i = E(y_i|a)$, $\phi$ is the dispersion parameter and $a(.)$, $b(.)$ and $c(.)$ are known functions.

For the inference based on the GLM to be valid we assume that the observations are statistically independent and the response variable in the data is linearly related to the covariates through the correctly specified link function. Furthermore we assume that the variance is no longer constant but is a known function of the mean. However in some situations such as when the underlying distribution is a Poisson model there may be over dispersion in the variance even for a simple random sample. There are statistical techniques that can be implemented to solve problems such as these, see for example Breslow (1996).

In general the GLMM are relatively difficult to fit for SAE, so are often avoided. Although the variable of interest is not always linearly related to the predictors, a GLMM may not be fitted, for example when modelling poverty a LMM is fitted to the log-transformed expenditure and then once predictions are made they are transformed into a binary indicator of poverty. An example of a GLMM being fitted is in the case of modelling prevalence of diarrhoea as

in Haslett et al. (2014a). The diagnostics presented in this thesis are not directly trialled on GLMMs used in SAE, however there is no reason to believe the methods cannot be used on a GLMM and this is a future area that could be researched.

## 2.3 Model Fitting

When fitting a model within the SAE framework, there are several factors that need to be taken into account. These include the structure of the data, as when the model is based on a linear mixed model or a generalized linear mixed model the variance components need to be estimated which can be problematic. Furthermore the data is usually collected from a sample survey using sampling methods such as stratification, clustering and weighting, therefore it is imperative the sample design is taken into account. Working with complex sample design (stratification, clustering weighting) leads to further issues of estimating the variance and uncertainty of the estimates.

### 2.3.1 Variance Components

The use of mixed models in small area estimation leads to the corresponding issue of correctly estimating the variance components of random effects. There are several possible methods, with some being more appropriate than others.

For simple random samples, Henderson's equation (Henderson, 1953) gives the best linear unbiased estimate (BLUE) and the best linear unbiased predictor (BLUP) even for non-normal data. Henderson (1953) proposed three different methods where these are estimated essentially by a reduction in the sum of squares. The first method is the simplest but is only applicable when there are random components, the second allows for fixed and random effects and the

third approach allows for fixed and random components as well as correlated data. The third approach can be computationally expensive when the number of components (e.g. the small areas) increases. The first two approaches use the ANOVA sum of squares, with the third approach using reduction in sum of squares through fitting constants. Henderson's equation gives the best linear unbiased estimator (BLUE) and the best linear unbiased predictor (BLUP) even for non-normal data. There is an extensive outline in Henderson (1953).

The minimum norm quadratic unbiased estimation (MINQUE) was proposed and is described in Rao (1972) and does not need the assumption of normality. The ANOVA method can also be used, however variance estimators using this method are inefficient when the data is unbalanced (Jiang, 2010).

The maximum likelihood estimation (MLE) and restricted maximum likelihood (REML) methods have grown in popularity in recent years. These methods can be used to estimate both fixed parameters and variance components. The MLE process selects values for the model parameters that maximise some given likelihood function. For unbalanced designs an explicit numerical expression is not possible, instead iterative procedures are needed to obtain the likelihood estimates (Rao, 1997). Furthermore it does not properly consider the degrees of freedom when estimating the fixed components in the model and results in underestimation of the variance components; this continues to increase as the number of fixed parameters increases (Rao, 1997). In real life scenarios the normality assumption is unlikely to hold, therefore the quasi-likelihood approach can be used for deriving the REML estimator (Jiang, 2010). An outline of the computational steps required to use the various methods is given in Searle et al. (2009). These estimation techniques can all be performed computationally, however some are more computationally extensive than others.

Moving beyond simple random samples, estimating variance components from sample surveys especially ones not designed for SAE can be a complex problem. In sampling schemes,

applied in the developing world, there are rarely many small areas with more than one primary sampling unit (psu) which makes it difficult to estimate both small area-level and psu-level variance components. This situation is explained in the next section.

### 2.3.2 Complex Survey Data

The collection of the survey data $Y$ used in SAE is commonly via a complex sampling design, consequently it is imperative the sampling design is incorporated into the first stage regression model. If the selection probability is disregarded, the regression parameter for the fitted model will be biased and the corresponding small area estimates and standard error estimates will be biased. Neglecting the sampling probability would assume that the sampling plan is based on simple random sampling (SRS). This is the most basic form of sampling and is rarely used in practice: it is more of a logical starting point and used as a theoretical underpinning for more complex methods (Lehtonen and Veijanez, 2009). In SRS each member of the population has an equal probability of selection, so SRS is an example of a self-weighting sample. In SAE there are usually several sampling techniques used. The methods applicable to the examples used in the remainder of the thesis are outlined below.

Stratified sampling is when the population is separated into non-overlapping mutually exclusive groups called strata, with observations in the same stratum being similar to each other. From here, a SRS of observations are sampled from every group. When the strata are relatively homogeneous, the variance of the sample estimator decreases. Lohr (1999) outlines several advantages of using stratified sampling: it protects from the possibility of obtaining a non-representative sample, it may be more convenient to run and it often gives more precise estimates.

Cluster sampling is a cost effective method and is usually done purely for administration

reasons (Lehtonen and Veijanez, 2009). A cluster sample is collected when the population is divided into $m$ different subgroups, ideally these subgroups will be representative of the population. A predefined number of clusters are sampled and then elements in that cluster are sampled. The estimators, which are generated from the cluster sample, tend to have a larger variance compared to SRS.

Systematic sampling occurs when one randomly selects a sampling unit within the first $k$ units and then every $k^{th}$ individual is sampled from the population. The value of $k$ is determined by the desired sample size.

In practice, several of these methods are combined in a sample design, this is known as multistage sampling. Due to the differing selection probabilities for units, it is important that sampling weights are available to researchers and employed in the analysis. For example, the Cambodian Socio Economic Survey (CSES), which was collected in 2009, used stratified and cluster sampling, followed by systematic sampling within the chosen clusters. The Nepalese Demographic Health Survey used a slightly different sampling method. They separated Nepal's thirteen domains into urban and rural, with exception of one particular domain where there was no urban area. From here a two-stage sampling method was used within each stratum, where the first stage selected enumeration areas (EAs) based on probability proportional to size; these are defined as the clusters. A ratio of approximately 1:2 (urban to rural) EAs were selected, resulting in the selection of 289 EAs. The second stage resulted in a set number of households being selected from each EA.

#### 2.3.2.1 Informativeness and Analytic Inference using Complex Surveys

For complex surveys, it is imperative in a design based context that sampling design and weighting is incorporated. In a model based context, selection probabilities instead need to be incor-

porated into the model as auxiliary variables. In both cases, adjustments are needed for any non-response or informativeness. Informativeness is when then sampling weights are related to the values of the model outcome even after conditioning on model covariates. This occurs when the survey sampling design and selection probabilities are correlated with the variable of interest even when conditioned on the explanatory variables. In this situation the observed outcomes are no longer representative of the population outcome. Also if there is non-response this can cause issues. Informativeness often arises when there is differential non-response across subgroups in the survey and where responders even within subgroup differ from non-responders in the same group. However in the case of poverty mapping in most developing countries, informativeness due to differential response rate is seldom an issue because response rates for household surveys usually have very high response rates, often in excess of 90% (UN Department of Economic and Social Affairs Statistics Division, 2005, ch. 12, sec. 29, p. 502). Where there is a very high response rate, differential non-response in not an issue, and so informativeness is also not an issue, because selection probabilities and missing subgroup information is not an issue. Pfeffermann and Sverchkov (2009) and Skinner and Wakefield (2017) outline how to make inference under complex sampling design and informative sampling.

### 2.3.3 Variance Estimation

When working with complex sample survey data or non-linear statistics, simple formulae cannot be used to compute the variance of an estimator, rather alternative methods are employed. There are several techniques that can be used. One is Taylor linearization (for continuous and differentiable variables), also known as the delta method and is based on a second moment calculation from calculus. Taylor's Theorem allows linearization of a smooth non-linear function of the mean (Lohr, 2009). Then there are the resampling methods: balanced repeated replication (BRR), jackknife repeated replication (JRR) and bootstrapping. These methods treat the

sample as if it were a population itself (Lohr, 1999). All the methods can be used to estimate parameters from a complex sample survey or to estimate their variance.

#### 2.3.3.1 Taylor Linearization

The Taylor linearization is similar to the sandwich estimator (Lumley, 2004). It approximates some non-linear function, where the variance of the function is based on the Taylor series approximation of the function (Lee and Forthofer, 2006). This method provides approximate estimates for the variance of first order statistics (Kish and Frankel, 1974).

Taylor's linearization has the advantage of being incorporated into many statistical software packages to estimate the variance of non-linear function and is the default option in most of these e.g. R, STATA and SAS.

There are several disadvantages to the method, such as the sample taken can influence the accuracy of the estimates and when the sample size is small, the variance may be underestimated (Lohr, 1999). Additionally, partial derivatives are also needed when complex functions using weights are used (Lohr, 1999). Furthermore it is difficult to apply Taylor's linearisation method to statistics that cannot be expressed as function of population totals or the mean or non-smooth statistics such as medians (Lohr, 2009)

#### 2.3.3.2 Balanced Repeated Replication

Balanced repeated replication (BRR) is designed for surveys with exactly two PSU's in each stratum (however these can contain pseudo samples). A half sample is taken by deleting one PSU from each stratum. This is then repeated a large number of times and the target statistic is calculated based on the data from the half samples. This method has the advantage of being less computationally intensive compared to other methods such as the bootstrap (described in

subsection 2.3.3.4). Lohr (2009) outlines the variance estimator using BRR for smooth non-linear population totals is asymptotically equivalent to the linearization method, but is also able to estimate the variance of non-continuous functions like quantiles, unlike Taylor's theorem. Lohr (1999) explains that BRR calculates the variance by assuming the sampling is done with replacement which is seldom the case, therefore it is likely to overestimate the variance when a sample is taken without replacement though the effect is usually very small. Lee and Forthofer (2006) also note that when the estimator is non-linear, the estimate of variance is slightly biased.

#### 2.3.3.3 Jackknife Repeated Replication

The JRR refers to the second-order estimation motivated by jackknife estimation (Kish and Frankel, 1974). Unlike the BRR, it can be used when a multistage sample has been conducted and can be applied to estimators that can not be expressed in terms of a formula (Lee and Forthofer, 2006). The JRR was first proposed by Quenouille (1949) as a non-parametric technique to estimate bias. However it was Tukey (1958) who used it for variance estimation. It uses a replicated resampling method where it removes one subsample at a time from the parent sample (the full sample) and generates the pseudo estimator based on this sample. The variance of the estimate is approximated using the pseudo estimates. Even if the point estimate is complex, the jackknife variance estimator will be approximately correct if the sample size is large (Lee, 2008). However, multistage sampling makes the jackknife method for variance estimation complicated and like Taylor's method it is usually assumed that the PSUs are sampled with replacement, which is seldom the case.

#### 2.3.3.4 Bootstrapping

As with the other replication methods the sample is treated as a population and samples are drawn from this population. A sample of size $n$ is taken with replacement from the original sample. Although this new sample is of the same size, it will be different from the original sample due to sampling with replacement; this means some observations will be included more than once, whereas some of the observations will be omitted. After a sufficient number of resamples are taken, where each sample is known as a "bootstrap", the mean of the overall test statistic is calculated, along with the variance. The bootstrap can be parametric or non-parametric. For example, Rao and Molina (2016) use parametric bootstrap to predict the $y$ in non-sampled areas. If the empirical probability mass function of the samples is similar to the probability mass function of the population then the corresponding samples which are generated should behave like samples taken from the population (Lohr, 1999). The variance of the estimator will be influenced by the number of bootstrap samples taken, therefore in order to reduce the computational error a large number of bootstrap estimates are usually taken, however this can make the process very computationally intensive. Often in small area estimation examples there is complex survey data being used. For complex samples with stratification and clustering, it is the clusters that are sampled rather than the secondary sampling. Rao (2003) provides an example of how to deal with the situation of using complex data.

## 2.4 Small Area Estimates

To generate the small area estimates from a fitted model there are several model based approaches that can be used. These consist of both frequentist and Bayesian methods. The first three are extensively outlined in Ghosh and Rao (1994), Rao (2003) and Rao and Molina (2015),

with the frequentist methods being the EBLUP and the Bayesian method being the EB and HB. The fourth method is an outlier robust method developed by Chambers and Tzavidis (2006).

### 2.4.1 EBLUP

The empirical best linear unbiased predictor (EBLUP) is a frequentist method of small area estimation used for mixed models. There are two main steps in obtaining the EBLUP. Step one involves deriving the best linear unbiased predictor (BLUP), which minimises the mean square error in the class of linear unbiased models; however this depends on knowing the variance and covariance of random effects. Secondly once the BLUP formula is obtained, estimates of the variance and the covariance are inserted. As these are unknown in most applications, they can be estimated using methods such as ML or REML (outlined in Section 1.3.1). By replacing the variance components with the estimates, the empirical BLUP (EBLUP) is generated. This process can sometimes be iterative. Using (2.3) and assuming $(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$ is an estimate of $(\sigma_v^2, \sigma_e^2)$ then the small area estimates from the EBLUP model can be shown as:

$$\hat{\theta}_i = \bar{\mathbf{x}}_i'\hat{\beta} + \hat{\gamma}(\bar{y}_i - \bar{\mathbf{x}}_i'\hat{\beta}) \tag{2.5}$$

where $\hat{\gamma} = (\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i)^{-1}\hat{\sigma}_v^2$, and $\hat{\beta} = (\mathbf{x}'\hat{\mathbf{V}}^{-1}\mathbf{x})^{-1}\mathbf{x}'\hat{\mathbf{V}}^{-1}\mathbf{y}$, where $\hat{\mathbf{V}}$ is the estimator of $\mathbf{V} = \mathbf{R} + \mathbf{ZGZ}'$ and $\mathbf{R}$ and $\mathbf{G}$ are both diagonal matrices, with $\mathbf{G}$ containing variance components and $\mathbf{R}$ being the covariance of the error terms.

### 2.4.2 Empirical Bayes

The Empirical Bayes (EB) and Hierarchical Bayes (HB) methods are much more broadly applicable than the EBLUP, these can be used to model data that is non-continuous such as for binary

or count data, and for models that are not regression-type models. When the model is normal and linear the EB method will produce the same estimators as the EBLUP (Rao and Molina, 2015). The reason that, for a LMM, it is sometimes said that EB and EBLUP are similar or even identical is that priors for the covariance structure in EB are usually chosen in the same way as for EBLUP. However even if the regression parameter estimates are identical, the MSE estimates for EB and EBLUP may not be identical, this depends on whether allowance is made for the estimation of the variance of the errors in the LMM.

The EB could be thought of as an approximation to the fully Bayesian method as it uses prior and posterior distributions like the Bayesian approach. However the density of the small area prediction is considered as part of the postulated model and can be estimated from the data. Rao and Molina (2015) outlines the optimal estimator of the value of $\theta_i$ is given by its conditional expectation given $\hat{\theta}_i$, $\beta$ and $\sigma_v^2$:

$$E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\mathbf{x}_i'\beta \tag{2.6}$$

where $\beta$ and $\sigma_v^2$ can be estimated from the marginal distribution given by $\hat{\theta}_i$ using the maximum likelihood or restricted maximum likelihood and substituting $\hat{\beta}$ for $\beta$ and $\hat{\sigma}_v^2$ for $\sigma_v^2$, this consequently gives the EB estimator of

$$\hat{\theta}_i = \hat{\gamma}_i\hat{\theta}_i + (1-\hat{\gamma}_i)\mathbf{x}_i'\hat{\beta}. \tag{2.7}$$

### 2.4.3 Hierarchical Bayes

The hierarchical Bayes (HB) method uses Bayes theorem to gain the posterior mean of the model parameters. This is similar to the EB approach where it uses subjective priors for ($\lambda$)

(based on existing knowledge) or more commonly diffuse or non-informative priors (as subjective priors are rarely available in SAE) to arrive at the posterior distribution, whereas EB substitutes suitable estimates of model parameters.

In HB the posterior density $f(\mu|y)$ of the small area estimates $\mu$ is obtained by combining the prior for $\lambda$ (where these are either diffuse or subjective) with the conditional model of $y$ given $\mu$ and the posterior density of $\mu$; from here Bayes theorem is used to gain the density of the $\mu$ (Rao, 2003). From this distribution, the posterior mean and variance can be obtained to get the small area estimate and the uncertainty surrounding the estimate.

The HB may use a Markov chain Monte Carlo (MCMC) method to draw the samples for the model parameters. There are several resampling methods, which can be used such as Markov Chain, Gibbs Sampler or the Metropolis-Hastings algorithm. When using the HB, it is important to consider whether the diffuse prior can lead to improper posteriors. This is explained further in Rao (2003) and Rao and Molina (2015).

### 2.4.4 M-Quantile

The previous methods of small area estimation use mixed effect models, with the random effects accounting for the differences between the small areas. Furthermore the previous methods often required distributional assumptions. Chambers and Tzavidis (2006) introduced M-quantiles for small area estimation. This method is robust against departure from distributional assumptions as it does not require formal specification of the random effect distribution or a hierarchical structure, making it robust against outliers. The method is based on modelling quantile like parameters on the conditional distribution of the variable of interest, rather than specifying the random effects.

Like the EB method, the M-quantile method separates the data into the in sample and

out of sample data. The survey data is then used to fit an M-quantile model to gain model parameters $\vartheta_i$ and $\beta_\psi(\vartheta_i)$, where $\vartheta_i$ is the quantile of interest in small area $i$ and $\psi$ is an asymmetric influence function. Using the estimated model parameters, an estimate is generated for each member of the non-sampled population using

$$\hat{y}_k = x_k'\hat{\beta}_\psi(\hat{\vartheta}_i) + e_k \quad k \in r_i \tag{2.8}$$

where $e_k$ is generated from the empirical distribution of the model residuals fitted to the survey data. The regression coefficient for each quantile can be estimated via iterative weighted least square analysis.

The $m_i$ quantile of small area $i$ can then be approximately estimated by

$$\hat{m}_i = N_i^{-1}\left\{\sum_{k\in s_i} y_k + \sum_{k\in r_i}\hat{y}_k\right\} \tag{2.9}$$

where $N_i$ is the population size in area $i$. The first part of the equation consists of the observed values $y_k$ for the $k$ sampled units, denoted $s_i$, within the area. The second component consists of the predicted values $\hat{y}_k$ of the $(N_i - n_i)$ non-sampled units denoted $r_i$.

The non-sampled component $\hat{y}_k$ in (2.8) is then repeated a large number of times, each time $e_k$ is drawn from an empirical distribution. This is combined with the sampled units and averaged over the number of repetitions to gain the M-quantile small area estimates.

Chambers and Tzavidis (2006) outline the advantages of using the M-quantile model for small area estimation. However the conditional distribution between the explanatory and response data needs to be well defined, when the response is nominal or multivariate, there may be difficulty using the M-quantile method, as there is not a logical method to order the response.

## 2.5 Diagnostics for Small Area Estimation

Regression diagnostics are used to check that the model assumptions hold, as well as identifying the influence that an individual or subset of data have on the outcome of interest. There are a range of diagnostic techniques for linear models such as checking for any patterns in the residuals, to ensure a linear model holds. The residuals or standardized residuals are plotted against the model fits to ensure linearity and homogeneity in the variance. Furthermore normality is tested through q-q plots as well as tests such as the Anderson-Darling test.

Checking for influential observations is also very important in order to identify if any observation(s) are having a significant impact on the regression, (although the central issue in this thesis is whether outliers affect the small area estimates). Cook and Weisberg (1982) define statistical influence analysis as assessing the effect small data perturbations have on the estimated regression parameters. A data point being highly influential is not necessarily problematic, but it can be a useful indicator of any observations that are having a significant impact. Measures of influence include the Cook's distance, this measures the overall effect of deleting an observation on the fit of the model to the observed data; the DFBETA is a scaled measure of the changes in each of the model parameters when a given observation is removed, and DFFITS measure the change in a fitted value for an observation when that observation is deleted, this being similar to the studentised residuals. Belsley et al. (1980) and Cook and Weisberg (1982) cover influence diagnostics for linear models in more detail. These methods have been adapted for survey data to incorporate complex sampling structure, see for example Li (2007), Preisser et al. (2008), Li and Valliant (2009) and Valliant (2010). At the model fitting stage, small area estimation can include linear mixed models (LMM), where Demidenko and Stukel (2004) outline a method to diagnose the observations that are influential. Additionally case-deletion diagnostics are used to identify influential subjects and observations in LMM in Pan et al. (2014). Although the

influence methods have been adapted to account for survey design and mixed models, there has been little work on adapting an influence diagnostic for small area estimation. This may be because even for linear models in SAE, the diagnostic focus is slightly different to the diagnostics for ordinary least squares (OLS) linear models or LMM. For SAE we are less concerned about how the regression parameters are influenced by the data, the focus instead being on how the predicted responses when aggregated to the small area level are affected by the model and the data.

In general, regression models are fitted in order to make predictions for future, unspecified data. However in SAE, the purpose of the model is very specific, with predictions required for several known sets of covariates, corresponding to the characteristics in small areas. Diagnostics in general tend to focus on the 'training data' to which the model is fitted. Brown et al. (2001) outline diagnostics that can be used to evaluate small area estimation methods, these include checking that the expected model parameters explain a significant proportion of the variation in the small area estimates. Some of the other diagnostics are focused on ensuring that the model based small area estimate should be approximately consistent with the expected value of the direct estimate, the model based estimate having a low mean squared error and importantly the small area estimate being able to inform the user. The model can be checked using standard model diagnostics; such as residual diagnostics, which are outlined as internal checking procedures by Rao (2003).

Brown et al. (2001) expand on four specific diagnostics for SAE, these being: bias diagnostics, a goodness of fit diagnostic, a coverage diagnostic and a calibration diagnostic.
Bias Diagnostics are used to ensure that estimates are unbiased. SAE is primarily used when direct estimation does not provide reliable estimates, however the direct estimates can be used to evaluate the goodness of fit of the SAEs. This can be tested by plotting the small area estimates against the direct estimates, or some transformation of both the direct and model based

estimates. Haslett et al. (2013) and other similar poverty mapping examples look at the standardized difference between the direct and estimated poverty levels at high levels of aggregation; anything with a standardised absolute difference less than two could be deemed acceptable and values larger than this would be flagged. The coefficient of variation (CV) is sometimes used to assess the reliability of the estimates, however this should be used with caution as in many cases the sample may be too small in the area and provide unreliable estimates. Therefore it is important to aggregate up to a sufficient level where the sample size is large enough to provide reliable estimates, for example in poverty mapping applications the sample sizes in the small areas are typically very small or zero.

The goodness of fit diagnostic is used by comparing the model estimate to the direct estimate by inversely weighting the squared difference by the variance and summing over all areas. This is tested against a $\chi^2$ distribution.

The coverage diagnostic compares the 95% confidence interval for the direct estimate with the model based estimate and measures how many times they actually overlap. Assessing the proportion of time the confidence intervals overlap can help determine if small area random effects need to be included.

Calibration can be used to assess how much the modelled estimates differ from the direct estimates when aggregated up to high levels of aggregation. Although this might be used when it is obvious that some large areas have differing model based and direct estimates, bench marking it is more often important to ensure aggregate SAE models match published survey estimates, as illustrated in Haslett et al. (2014b).

Furthermore, when using Bayesian techniques as in HB there are several additional aspects that need to be checked such as convergence and run length. Molina et al. (2014) also consider a validation diagnostic to ensure that the assumed model fits the data. The standard-

ised cross validation residual is considered, which looks at the predictive distribution of each observation when that observation has been deleted from the sample.

When it is identified that there are outliers or errors in the covariates, adjustments can be made in the model fitting procedure. Pfeffermann (2013) outlines recent developments in accounting for measurement errors in covariates, as well as treatments for outliers, and Chambers et al. (2014) and Arima et al. (2016) focus on making SAE robust to model outliers that are present in the training data. Current methods tend to apply only to the relationship between the training data and the fitted model, rather than the final small area predictions. Despite rapid advancements in small area model fitting techniques, diagnostics checking the validity of these models and identifying outlying or unusual small area estimates (rather than model outliers) have largely been neglected. When the estimate for a particular small area is felt to be unusual, for example based on expert opinion, it would be useful to explore which variables and observations appear to be the cause, so that possible remedies can be sought.

Previous work has focused largely on estimating and reducing the mean squared error (MSE) or diagnostics using the training data to fit the model, whereas this thesis focuses on diagnostics that apply to the final small area estimates.

# Chapter 3

# Small Area Estimation and Poverty

## 3.1 Introduction

Nearly half the world's population live on less than $2.50 a day, with more than 1.3 billion living in extreme poverty of less than $1.25 a day (United Nations Development Programme, 2014). Every year billions of dollars in Official Development Assistance, private aid and public aid are allocated to attempt to eradicate poverty. The use of targeting becomes important to ensure aid is distributed to the people who are experiencing the greatest deprivation. Poverty maps, which are a graphical method to show the concentration of poverty within a country, have become a useful tool for aid distribution as they show a geographical representation of the different levels of deprivation in the country. This chapter will outline different measures of poverty that are used, as well as undernutrition measures. A section outlining poverty mapping follows this. Lastly, different model fitting techniques that can be applied to generate small area estimates for the particular focus of poverty estimation are outlined; these include the ELL method as well as several alternatives including Molina & Rao's Empirical Bayes method, Hierarchical Bayes and the M-quantile method.

### 3.1.1 Measures of Poverty

Poverty has many facets, including having a lack of resources, having a low income or being deprived of opportunities. It can be divided into two main classes; human poverty and income poverty. Human poverty is defined as the denial of choice or opportunity in life (UN, 1997), but this can be difficult to measure due to factors such as susceptibility to violence being difficult to measure. The Multidimensional Poverty Index (MPI) can be considered as a measure of human poverty; this takes into account the interactive harm of multiple deprivations, however it does not take into account factors such as access to credit or susceptibility to violence and consequently is not a perfect measure of standard of living. The MPI is of limited use for targeting and intervention as it is not obvious which aspect of deprivation is present. On the other hand income poverty measures the financial deprivation only. Coudouel et al. (2002) defines a household to be in poverty if they have inadequate resources to meet their needs. A poverty indicator can be used to measure this which is derived from the per capita income and expenditure of a household. The measure first introduced by Foster et al. (1984), henceforth referred to as the FGT, can be defined as

$$P_i^{\alpha} = N_i^{-1} \sum_{k=1}^{N_i} \left( \frac{z - E_{ik}}{z} \right)^{\alpha} \mathrm{I}(E_{ik} < z), \quad i = 1, ..., I \quad k = 1, ...N_i \tag{3.1}$$

where $E_{ik}$ is the measure of expenditure for the $k^{th}$ individual/household in small area $i$, $N_i$ is the population size in small area $i$, $z$ is the defined poverty line, I is an indicator variable that takes the value 1 if the expenditure is less than value of z and is zero otherwise. Finally, $\alpha$ is a measure of sensitivity. The headcount index is defined when $\alpha$ = 0, which measures the proportion of people who fall below the poverty line. Other possible values for $\alpha$ are 1 and 2, where these measure the average poverty gap and severity respectively.

The poverty line $(z)$ can be defined in several forms, for example organisations such as

the World Bank often defined people who live on less than $1.25 a day as being in extreme poverty and people living on less than $2 as being in poverty. In 2015, the global poverty line was updated from $1.25 to $1.90, this was to reflect the changes in the cost of basic needs (World Bank, 2015). These monetary values are adjusted from country to country to reflect purchasing power parity (PPP)[1]. The poverty line can also be defined in terms of the Cost of Basic Needs (CBN) (World Bank, 2015); this is determined by taking into account the cost an individual faces in order to buy enough food to consume 2100 calories each day as well as the non-food expenses, such as rent (Haughton and Khandker, 2001). Because the cost of living differs throughout the country, particularly between rural and urban areas, different poverty lines may exist throughout the country.

### 3.1.2 Undernutrition

Poverty can have a significant adverse effect on child development. Undernutrition is a leading cause of death in children in the developing world (WHO, 2014). It accounts for 45% of deaths in children under the age of five, with an estimated half of these deaths being preventable (Wang and Chen, 2012). Therefore, it is important to gain reliable estimates of undernutrition at finer levels of aggregation, as this allows policies and aid such as feeding supplements to be targeted to the children who are most at risk.

To generate the undernutrition rates in each small area, anthropometry measurements for children less than five years old are used based on their weight, height and age. There are three measurements used to assess a child's health. These include the standardized weight-for-age (WAZ), height-for-age (HAZ) and weight-for-height (WHZ). These measurements are standardized against an international reference population (World Health Organization and UNICEF,

[1] Where PPP is an economic theory that uses an index to compare different countries currency through assessing how many units of each currency it takes to purchase a basket of goods and services (Hall, 2018)

2009). A child is defined to be stunted if their HAZ is below -2 and they are severely stunted if HAZ is below -3. A child is underweight if their WAZ is less than -2 and severely underweight if it is less than -3 and similarly a child is defined as being wasted if WHZ is less than -2 (this corresponds to a weight to height ratio more than two standard deviations below the median of the reference population) and severely wasted if it is below -3. For now the focus will be on the WHZ measure.

Adapting (2.2), $WHZ$ can be modelled as

$$WHZ_{jkl} = \mathbf{x}'_{jkl}\beta + v_j + \eta_{jk} + e_{jkl} \tag{3.2}$$

where $WHZ_{jkl}$ is the score for the $l^{th}$ child in the $k^{th}$ household in the $j^{th}$ cluster, and $\mathbf{x}_{jkl}$ is a vector of auxiliary variables. The nested error is three-fold, accounting for the unexplained variation at cluster, household and child levels, using random effects $v_j$, $\eta_{jk}$ and $e_{jkl}$ respectively. An error term at small area level may also be required unless there are suitable contextual variables included in $x_{jkl}$ that are the same for all children for a given $j$ or subset of it, for example sub-cluster means.

When fitting the model to the survey data, the sampling structure can been taken into account as well as adjustments being made for the variance structure. Adapting the FGT in (3.1) the wasting rate $(\hat{W}_i)$ can be estimated for each small area as:

$$\hat{W}_i = N_i^{-1} \sum_{j,k,l \in i} \mathbf{I}(\widehat{WHZ}_{j,k.l} < -2.00) \tag{3.3}$$

where $N_i$ is the number of children under the age of five in small area $i$ and $\mathbf{I}$ is an indicator variable that receives a 1 if the $WHZ$ is less than -2.

### 3.1.3 Poverty Mapping

A poverty map is a geographical profile displaying the level and concentration of poverty within a country. It shows the spatial representation and analysis of human well being and poverty by combining micro level data such as individual and household level survey data, with macro level data which concerns the population as a whole. Poverty maps can be built using a range of data types including censuses, surveys and administrative data (World Bank, 2011a), and are commonly used to show the small area estimates of poverty. This can be used to provide an informative representation of the geographic distribution of poverty within a country (Hentschel et al., 2000). Geographic Information Systems (GISs) are then used to display the disaggregated information using geographic coordinates. These maps provide a useful tool for policy advisors and aid organizations to determine where to target policies and aid in order to alleviate poverty. Henninger and Snel (2002) have produced a report outlining the importance of poverty mapping and the increasing importance it has to assess social and economic problems.

There are various methods that are used to generate a poverty map, with various techniques being used over the past 30 years to determine poverty at finer levels of aggregation. Each of these methods have advantages and disadvantage, however one method which has been used extensively in the developing world is the ELL methodology (World Bank, 2011a); this will be discussed in more detail in Section 3.3.1. Davis (2003) produced a paper outlining these various methodologies, which have been used to produce poverty maps, briefly summarised below. Poverty mapping can be done via multivariate weighted basic-needs index; this involves a weighting scheme. Another technique uses principal components, a statistical technique to reduce a large set of variables by extracting a linear combination, which best describe the variable. This was commonly used in Mexico. However care needs to be taken when variables are negatively correlated (for example stunting and wasting in children under 5 in Nepal). A

further extension to this is principal components over time, which extends the analysis to include time. There is also factor analysis which is used to describe the relationship among many variables. This has been used on South Africa with the 1996 census (Davis, 2003). Another type of poverty mapping is combining qualitative information with secondary data; however this is generally used for mapping food security rather than poverty. If the census data collects information which directly relates to welfare then direct measurements of census data can be used to create a poverty map. Similarly, if the survey collects data from a sufficiently large number of people then direct estimates of survey data can be used. However in practice this is rarely feasible, due to either insufficient data relating to welfare and income of households being collected, or an insufficient sample size being used.

## 3.2 Small Area Estimation techniques for poverty

Chapter 2 outlined several commonly used model based estimation methods used to generate small area statistics; however these usually apply to some linear function such as means or totals, whereas poverty tends to be measured by more complex non-linear functions such as the FGT shown in (3.1). This section will outline several popular choices that can be used for the non-linear functions commonly used to measure poverty or inequality.

Survey data is commonly used to get information on income and expenditure, however the survey data alone is insufficient for estimation for small areas due to the limited sample size, meaning it is impossible to generate reliable estimates at fine levels of aggregation. A secondary data source such as a census is used to draw strength for the estimates. As a result of the census data supplementing the survey data, poverty can be predicted with a greater level of precision at aggregated levels. This allows more efficient use of aid allocation through a finer level of poverty targeting, for example reliable estimates at sub-district level (typically

15,000-30,000 people) are possible using the ELL methodology (Elbers et al., 2002). Molina and Rao (2010) and Molina et al. (2014) extended the EB and the HB respectively to allow for non-linear response functions, such as poverty measures. This version of the EB will be referred to as EB_MR. In practice the EB_MR and HB methods have mostly been applied to larger small areas in comparison to the ELL method. For example the ELL method often will have over 1000 small areas, for example Cambodia had 1621 small areas (Haslett et al., 2013), whereas Molina and Rao (2010) used 104 small areas when determining the poverty in Spain. Furthermore in Tzavidis et al. (2008) there were only 36 small areas and in Marchetti et al. (2012) there were only 10 small areas when using the M-quantile method. The ELL method uses unit record census data, when modelling continuous covariates. With the EB_MR and HB applications, unit record census data is not usually available and so proxy or model-based censuses are generated. This can be seen in Molina et al. (2014) where in the case of Spain there was no unit record census data available, so they replicated the matrix of covariates from a larger survey a number of times until they matched the weights for the survey. This was used as a proxy for the true census. This can only be done for categorical covariates.

An explanation of these four methods will be presented in the following sections.

## 3.3 Model Fitting Techniques

### 3.3.1 ELL

Elbers, Lanjouw and Lanjouw developed a methodology to formulate poverty maps. This method is sometimes referred to as the World Bank (WB) method or the ELL method; I will refer to it as the latter. Since 2003 the ELL method has been implemented in more than 50 countries to allocate billions of dollars of aid, some of these countries include Albania, Brazil,

Cambodia, Ecuador, Guatemala, Kenya, Mexico, Nepal and Uganda.

The initial step in the ELL method is to fit a nested model to the household expenditure $E_{jk}$ or rather the log expenditure $ln(E_{jk})$; as this helps to determine if a household has adequate money to live off and the log of this is taken due to the highly skewed distribution. A model is fitted using auxiliary information from the survey, census level means or administrative data. There is a critical assumption that all of the survey variables used in the model must also be contained in the census; they must be measured in the same way and have a similar distribution in both sources. The nested model can be written as:

$$Y_{jk} = E[\ln E_{jk}|x'_{jk}] + \varepsilon_{jk} = \mathbf{x}'_{jk}\beta + v_j + e_{jk} \quad j = 1,...,J \quad k = 1,...,N_j \tag{3.4}$$

$$v_j \sim N(0,\sigma_v^2) \quad e_{jk} \sim (0,\sigma_e^2)$$

where $\varepsilon_{jk}$ is the overall nested error that is decomposed into the cluster level error $v_j$ and the household level error $e_{jk}$, $j$ denotes the $j^{th}$ cluster as defined by the sample design and $k$ is the $k^{th}$ household within the cluster. The ELL model fits the random effects at cluster level as households within a cluster tend to be more similar to each other. Furthermore the cluster level and household level random effects are assumed to be independent and uncorrelated with the auxiliary variables. The relative importance of each of variance components is reflected in the ratio of $\sigma_v^2$ to $\sigma_e^2$. It is beneficial that a larger proportion of the variability is explained at the lower level of aggregation, consequently the unexplained variability in the cluster level errors should be small relative to the unexplained variability for the households within the clusters. This is because when the predictions are aggregated the household level variation tends to cancel out. The linear mixed model is used to take into account the multistage sampling method and more importantly the unexplained variation at various levels of aggregation. In order to ensure that $\hat{\beta}$ and $Var(\hat{\beta})$ are not biased the design of the sample survey is incorporated through

specialized statistical routines, see for example Haslett and Jones (2010), including weights to account for the stratification and a variance estimation technique that incorporates the clustering of the survey design (see Section 2.3.2). The alternative is model based and includes design variables in $\{\mathbf{x}_{jk}\}$.

After the model is fitted to the survey data, bootstrap estimates are generated for each member of the population, not just the sample. Bootstrapping is used to give the required joint distribution of the estimates allowing for the uncertainty about the model parameters. Often the fixed effects parameters are drawn from the multivariate parametric distribution $\beta^b \sim N(\hat{\beta}, V(\hat{\beta}))$, the cluster level effects $v_j^b$ are drawn randomly with replacement from the cluster level residuals. However, we only have cluster level residuals from the sample data rather than the census, as not all observations from the census are included. Lastly, the household level residuals $e_{jk}^b$ are drawn; it is usually assumed the household level errors are heteroscedastic (Elbers et al., 2003). In this case the household level errors are assumed to depend on a subset of auxiliary variables $Z$ where $g(\sigma_e^2) = Z\delta + r$ and $\delta$ is a vector of regression coefficients. Consequently, the predicted household level residuals depend on the auxiliary parameter estimates $\hat{\delta}$ and the variance matrix $V_\delta$. The resulting bootstrap estimate for a household in the population is

$$Y_{jk}^b = X_{jk}\beta^b + v_j^b + e_{jk}^b, \quad b = 1, ...., B. \tag{3.5}$$

These indicators are defined at household level, but the predictions are weighted by the number of individuals in the household to get person-level summaries. In the case of undernutrition models, the indicators are at child level The number of bootstraps taken tends to be around 100 in order to get reasonably precise estimates of the standard errors. If the number of bootstraps is small the estimates tend to be unstable. It is not the logarithm of expenditure which is of interest but whether a person is in poverty, so the $Y$ values need to be transformed as we require the expenditure per person, and the expenditure per person is compared to the poverty line. The

predicted expenditure then becomes $E^b_{jk} = \exp^{Y^b_{jk}}$. The expenditure can also be compared to the poverty line to gain the gap and severity.

This information is used to determine the bootstrap poverty incidence in each small area, by considering whether each household (and occupant) is above or below the poverty line. The bootstrap value of the poverty incidence for area $i$ is calculated by using (3.6)

$$P_i^{\alpha,b} = N_i^{-1} \sum_{jk \in i}^{N_i} \left( \frac{z - E^b_{jk}}{z} \right)^{\alpha} \mathrm{I}(E^b_{jk} < z). \tag{3.6}$$

The corresponding mean and uncertainty for each small area is then estimated by

$$\hat{P}_i^{ELL,\alpha} = B^{-1} \sum_{b=1}^{B} P_i^{\alpha,b} \quad and \quad MSE(P_i^{ELL,\alpha}) = B^{-1} \sum_{b=1}^{B} (P_i^{\alpha,b} - \hat{P}_i^{ELL,\alpha})^2. \tag{3.7}$$

The ELL method is not unique to income poverty but can be generalised to other measures such as expenditure poverty, and stunting, underweight and wasting in children.

### 3.3.2 Empirical Bayes

Molina and Rao (2010) extended the EB method to apply to non-linear functions such as poverty indicators; henceforth this method will be referred to as EB_MR. In the literature it has also been referred to as the empirical best prediction (EBP). This is outlined in full detail with a simulation study and application in Molina and Rao (2010), and a more concise explanation is given in Rao and Molina (2016). The following is a summarised explanation.

In EB_MR the total population can be separated into a vector of the sampled households $y_{is}$ and the non-sampled households $y_{ir}$ in each small area $i$. The two stage nested model is used to fit the sample data to the log of expenditure as in the ELL model, however it models the small

areas as random effects $u_i$ rather than clusters $v_j$ shown as:

$$Y_{ik} = \mathbf{x}'_{ik}\beta + u_i + e_{ik} \quad i = 1,...,I \quad k = 1,...,N_i \tag{3.8}$$

$$u_i \sim N(0, \sigma_u^2) \quad e_{ik} \sim (0, \sigma_e^2). \tag{3.9}$$

The best predictor in general (for any $\alpha$) can be estimated using Monte Carlo approximation by generating values $Y_{ik}^{(c)}$ for $c = 1,..,C$, for the non-sampled data, given the conditional distribution of the sampled data. The Monte Carlo approximation for the non-sampled data is shown as

$$\hat{P}_{ik}^{\alpha} = C^{-1} \sum_{c=1}^{C} f(Y_{ik}^{(c)})^{\alpha}, \quad k \in r_i \tag{3.10}$$

where $f(Y_{ik})$ is the required function of $Y_{ik}$ (in this case the FGT poverty estimator and $\alpha$ is the measure of sensitivity). The best estimator minimises the MSE and depends on the parameters $\beta$, $\sigma_u^2$ and $\sigma_e^2$. Using ML or REML, we estimate the model parameters and get the EB_MR; $\hat{P}_{ik}^{EB}$; for the non-sampled units. Usually all small areas will contain sample, so $\hat{u}_i$ is not necessarily zero; otherwise the estimates are synthetic.

Combining the estimates for the sampled and non-sampled data we gain the EB_MR estimate defined as:

$$\hat{P}_{ik}^{EB,\alpha} = N^{-1} \left( \sum_{k \in s_i} P_{ik}^{\alpha} \quad + \quad \sum_{k \in r_i} \hat{P}_{ik}^{\alpha,EB} \right) \tag{3.11}$$

where $s_i$ is the sampled data in small area $i$ and $r_i$ is the non-sampled units This method assumes knowledge of the linking between sampled units to the population. If this isn't possible, as usually the case, an adaptation can be made, however it is slightly less efficient. When there are no sampled units in the small area $i$, bootstrap estimates are derived for the small area level and

the individual level components $u_i$ and $e_{ik}$ where these are drawn from $N(0, \hat{\sigma}_u^2)$ and $N(0, \hat{\sigma}_e^2)$ respectively. These are synthetic estimates and essentially the same as the ELL if the random effects are defined in the same way. It is difficult to get an analytical approximation of the MSE, instead parametric bootstrapping is used, however this can be very computer intensive.

### 3.3.3 Hierarchical Bayes

Molina et al. (2014) developed a hierarchical Bayes (HB) method for non-linear functions such as poverty measures. This method has the benefit of being more computationally efficient compared to the EB due to not needing to perform parametric bootstrapping to estimate the MSE of the estimates. Rao and Molina (2016) give an overview of the method comparing it to both the ELL an the EB method. Like the HB outlined in Section 2.4.3 a prior distribution is placed on the model parameters and from here the posterior distribution is formed and a large number of samples are generated from this. If we have $d = 1...D$ samples, the posterior mean can be obtained by averaging over the drawn samples

$$\hat{P}_i^{HB,\alpha} \approx D^{-1} \sum_{d=1}^{D} P_i^{\alpha,(d)}. \tag{3.12}$$

Molina et al. (2014) considered a reparametrisation $P_i^{\alpha,(d)}$ based on expressing the model in terms of the intra-class correlation $\rho = \sigma_u^2/(\sigma_u^2 + \sigma^2)$. It is assumed that there is no informative sampling and therefore the population model fitted to the sample data holds for both the sampled and non-sampled units. With this reparameterisation along with non-informative priors, the Markov chain Monte Carlo (MCMC) sampling can be avoided and simulations can be pulled at random directly from the posterior distribution, where the parameters of interest are $\gamma = (u', \beta', \sigma_e^2, \rho)$. Using the initial sampled data $y_s$, $\rho$ is pulled from the joint posterior distribution, followed by each of the other model parameters. This process is repeated a large number of $D$

times. The model parameters $(\gamma^{(d)})$ are then used to determine the value of the non-sampled values $y_{ik}^{(d)}$. The full population is generated by combining the sampled vector $y_{is}$ with the non-sampled vector of estimates for $y_{ir}^{(d)}$, and from here the poverty estimate can be calculated.

### 3.3.4 M-Quantiles

Tzavidis et al. (2008) extended the M-quantile method to poverty estimation. Marchetti et al. (2012) outlined that the mean square error of the M-quantile estimate can be unstable when the sample size in each small area is small, and so they proposed non-parametric bootstrapping to estimate the MSE of M-quantile poverty indicators. This achieved more stable estimates than analytical techniques. This method is also outlined in Das (2016). The method involved resampling the model residuals and although it produces more stable estimates it is very computer intensive, for example Marchetti et al. (2012) noted it took 16 hours to run for a population of 1.4 million households, this is relatively small for poverty mapping projects.

### 3.3.5 Comparison of the SAE methods for poverty

There is a considerable amount of debate over which method is the best to use for small area estimation of poverty, however the availability and the structure of the data tends to influence which method is going to give the most reliable estimates. Each of the methods is based on different assumptions and if these are not met or the model is misspecified then the model will not provide reliable small area estimates. It is possible to do a simulation study to compare the methods, however in such cases the way data are generated can greatly influence the relative performance of the methods. Haslett (2016) and Rao and Molina (2016) have provided an outline and comparison of the ELL, EB_MR and HB methods and Haslett (2016) provides an extensive overview of the advantages and disadvantages of the ELL method. Das (2016)

compares the ELL, EB_MR and the M-quantile methods in more detail and provides a simulation study to compare which method performs the best. Furthermore, he found that the ELL method is computationally faster compared to the EB_MR and M-quantile. Another of the main differences between the ELL, EB_MR, HB and M-quantile is the specification of the random effects. The ELL specifies them at cluster level, whereas the EB_MR and HB account for the between area level effects and the M-quantiles method does not need to specify the distributional assumption of the random effects. One of the main arguments for the use of EB in favour of ELL is that it is claimed to be more efficient in terms of the MSE when there are strong area effects, however this is only significant when contextual area level effects are not incorporated into the ELL model. In developing countries, there tends to be a large number of clusters within a small area and random effects at cluster level tend to be more prominent than the area level random effects. The EB_MR and the M-quantile often specify matching the households in the survey with the census, however in practice this is seldom feasible.

This thesis is not focused on comparing which small area estimation technique is the best. Rather the focus is on diagnostic tools that can be applied in order to choose auxiliary variables, identify any anomalies in the small area estimates, and examine the adequacy of the estimates' precision. From here onwards the application of the examples will be on data which have used the ELL method, however the diagnostic techniques will also be applicable to the other small area estimation methods.

# Chapter 4

# Data

This chapter outlines the data that will be analysed in the following two chapters. The first example applies diagnostic techniques to small area poverty estimates from Cambodia; this data is outlined in Section 4.1. This section introduces some background information of Cambodia and then is divided into three main subsections: these being an outline of the 2010 census, an outline of the survey data and finally the model formulation used to produce the small area estimates of poverty in Cambodia. Section 4.2 outlines the second set of data analysed in small area estimation for wasting rates of children in Nepal. The section provides information on the Nepalese 2011 census, the Nepal Demographic Health survey (DHS) and the model formulation used to produce wasting estimates is described.

## 4.1 Cambodia

Cambodia is a developing country located in South East Asia with a total area of 181,035 $km^2$. It borders the Gulf of Thailand, Thailand, Laos and Vietnam. Cambodia has lagged behind in economic growth and development, where 28.3% of its population were living on

less than $1.25 (PPP) [1] per day in 2011 (UNDP, 2011). Furthermore, it was ranked 139th out of 187 in the human development index (UNDP, 2011). The high poverty rate and low human development can be partly attributed to years of suffering, conflict, civil war and corruption. Between 1975 and 1979, the rule by Khmer Rouge regime resulted in an estimated two million people dying due to execution, exhaustion or disease (Fletcher, 2009). The genocide devastated the country and left long term problems, such as political instability, low human capital levels and poverty. The high incidence of poverty has led international organisations, non-government organisations and the local government to make poverty eradication a priority in Cambodia.

As a result of targeting and other such development policies, poverty in Cambodia has decreased from 47% in 1993 to 30% in 2007 (Ibp, 2011), however inequality increased over the same period. Having reliable estimates of poverty at finer geographical areas in Cambodia means aid can be targeted to the poorest regions and therefore poverty eradication can be achieved in a more efficient time frame. The Cambodian Socio-Economic survey from 2009 was used to formulate a model to explain log expenditure and this was combined with the 2008 census in order to make poverty predictions.

### 4.1.1 Cambodian Census 2008

The second most recent Cambodian census was conducted on March 3rd 2008, with the financial and technical support of agencies such as the United Nations Population Fund, the Japanese International Cooperation Agency, the German Government and the Japanese Government. The month of March was chosen as this was deemed to be a time with a stable population and little international travel, and therefore a representative indication of the population structure could be obtained. This was the second census to be conducted after a 36 year gap, the first one being

[1] Purchasing power parity

in 1998. In the years preceding this, there was no census able to be collected due to the political instability and conflict. Following this period, a legislation was passed that a census must be carried out at least once every 10 years (RamaRao, 2008); this has the purpose of helping the Cambodian government keep up to date records of Cambodia's population characteristics.

The 2008 census contained three main forms. The first related to the enumeration of buildings; the second form asked questions with respect to household ownership, utilities and appliances; and the third form gained information on individual demographic indicators.

Census enumeration area (EA) maps were created to divide the country into different geographical locations. The EA maps were created in 2006 using Global Positioning Systems (GPS). The census included people within an EA who were in a household, or an institution, or homeless, or transient population. It did not include tourists, temporary visitors, refugees or foreign diplomats. Based on this definition of the population, the UN reported the population for Cambodia on census night to be 13.4 million people and 2.5 million households. 19.4% of the population lived in urban areas and the remaining 80.6% lived in rural areas. For administrative reasons the country was divided into a hierarchy of structural units. The largest unit area unit is region, this is followed by district, commune then village. An outline of the structure at various levels is given in Table 4.1, showing the mean number and minimum number of households and primary sampling units in each area level.

Table 4.1: Structure of the Cambodian census.

| | Province | District | Commune | Village | EA |
|---|---|---|---|---|---|
| Contains | 24 | 193 | 1621 | 14073 | 28455 |
| Mean Household | 117228 | 14578 | 1736 | 200 | 99 |
| Min Household | 7193 | 850 | 60 | 3 | 3 |
| Mean PSU | 1186 | 147 | 17.6 | 2 | |
| Min PSU | 66 | 10 | 2 | 1 | |

### 4.1.2 Cambodian Socio-Economic Survey 2009

The second source of data for the model is the Cambodian Socio-Economic Survey (CSES). This a nation-wide survey carried out by the Cambodian National Institute of Statistics. It was introduced in 1993 and since 2007 it has been conducted annually. The CSES is based on the Living Standards Measurement Study conducted by the World Bank (World Bank, 2011b). The main objective is to collect statistical information on the standard of living of the population. The survey contains five main sections. The first two relate to facilities of business behaviours in the village and the remaining three sections relate to household matters such as questions regarding income and spending, labour force participation and conditions of living. The combination of the forms helps estimate the economic conditions and the level of poverty for the country.

The 2009 CSES was carried out between January and December, with 1000 households being surveyed each month, giving a total of 12,000 households surveyed. There was a nearly 100% response rate with a total sample size of 11,971. The CSES was collected via multi-stage random sampling which included stratified, cluster and systematic sampling techniques. The villages were classified as the primary sampling unit (PSU); however the larger villages were divided up based on census EA. The PSUs were selected with probability proportional to size (PPS) within strata, where the strata are provinces divided into urban and rural. In total 720 PSUs were sampled: 240 in urban areas, and 480 in rural areas. Systematic sampling was then used to select 10 households from urban PSUs and 20 households from rural PSUs. Table 4.2 outlines the structure of the CSES and census showing the number of units at each level of aggregation.

Table 4.2: Structure of the CSES2009.

| | Province | District | Commune | Village |
|---|---|---|---|---|
| Contains | 24 | 171 | 621 | 715 |
| Mean Households | 499 | 70 | 19.3 | 16.7 |
| Min Households | 39 | 19 | 9 | 8 |
| Mean PSUs | 30 | 4.2 | 1.2 | 1.01 |
| Min PSUs | 3 | 1 | 1 | 1 |

### 4.1.3 Poverty Model for Cambodian SAE

The CSES data, GIS data and EA level census means were used to fit a model explaining log expenditure, which was then used to estimate the poverty rate. The following is a description of the method that was performed by Haslett et al. (2013). The CSES variables were restricted to variables which were measured and defined in the same way as they were in the census. In total, there were 36 possible variables from the CSES and 52 census means or GIS variables that could have been used as possible predictors. The variables consisted of numerical ones such as the household size, and categorical ones such as roofing type of the household. With numerical variables possible transformations were also investigated, and categorical variables were made into binary indicators. Taking into account the large number of variables as well as all the possible interactions, there were a large number of models to consider; this meant the preliminary model selection was largely automated using a process such as stepwise regression or best subset. Generally one model was fitted for the entire country, however regional effects were considered and region:variable interactions were included if found to be significant. Some applications of small area estimation fit separate models for each of the different strata, however this was not used in Cambodia so as to avoid over fitting. The selected model contained 35 variables. These are defined in Table 4.5 and the fitted model is shown in Table 4.6 in the Appendix. The variables in the model that end in "_*e*", are enumeration area (EA) level means, where EAs were the primary sampling units (also known as the cluster level means). Of the

35 variables there were four variables included as interaction terms, where these are denoted with "XS3". These variables behave differently in region 3, which is the capital city Phnom Penh, and so the variables were included to account for the different effect. The final model explains 65.6% of the variability in the natural log of expenditure. A greater proportion of the residual variance occurred at the household level rather than the cluster level where the ratio of the cluster to total residual variation is 0.267. This means that in general the between cluster variation for the log of expenditure was fairly well explained by the contextual variables.

The full ELL process, which was explained in Section 3.3.1, was used to generate the small area poverty estimates. For a more detailed report as well as to see the final small area estimates see Haslett et al. (2013).

## 4.2 Nepal

Nepal is a land locked country in South East Asia located between China and India. It has a total land area of 147,181 $km^2$, and is most well known for Mount Everest and trekking in the Himalayas. It is classified as a Hindu nation with 81.3% of the population reporting Hinduism as their religion.

Nepal is underdeveloped which has led to children being impoverished. In 2010, Nepal was reported as the poorest country in South East Asia and the fifteenth poorest country in the world (International Development Committee, 2010). Additionally, it had a high incidence of child undernutrition; in 2001 it had reported rates of 50.5% of children under five being moderately stunted, 48.3% being wasted and 9.6% being moderately underweight (UNICEF, 2006); stunting, wasting and underweight are defined in chapter two. The high incidence of child poverty and adverse outcomes to their health in Nepal makes small area estimation vital so food and funding can be targeted to the most vulnerable.

Table 4.3: Structure of Nepalese Census.

| | Region | District | Ilaka | VDC/Mun | Ward |
|---|---|---|---|---|---|
| Contains | 5 | 75 | 976 | 3973 | 36041 |
| Mean Children | 513226 | 34215 | 2627 | 645 | 71 |
| Min Children | 296508 | 376 | 9 | 4 | 1 |
| Mean Household | 384246 | 24616 | 1966 | 483 | 53 |
| Min Household | 212830 | 326 | 9 | 3 | 1 |
| Mean ea | 8115 | 541 | 42 | 10 | 1.1 |
| Min ea | 3761 | 86 | 4 | 3 | 1 |

### 4.2.1 Nepal Population and Housing Census 2011 (NPHC 2011)

The 2011 Nepalese population and household census was held on June 11th. It was the eleventh census conducted in Nepal, with the first being held in 1911. The population census collects information on all residents in Nepal at their usual place of residence. The homeless or mobile population were counted at the location they were traced to the last day of enumeration. The recorded population on the night of the census was 26,494,504 with 5,427,302 households. Of those 5,423,297 households were classed as residential properties and included in the census.

Nepal consists of three ecological zones (mountains, plains and terai), and five development regions (eastern, central, western, mid western and far western). Combining these together there were 15 domains, however the western development regions in the mountains were combined leading to only 13 domains. Nepal further consists of 75 districts and each district is further divided into EAs which consist of village development committees (VDC) in rural areas and municipalities in urban areas. In the case of Nepal, the small area level is at ilaka level.

The focus in this study is information on children under the age of five. Table 4.3 shows the structure of the households as well as children under the age of five at various levels of disaggregation. In total, this included approximately 2.5 million children. Of the households with children, approximately 27.5% had two or more children under the age of five.

### 4.2.2 Nepal Demographic Health Survey 2010

The Nepal demographic health survey (DHS) is a nationwide survey undertaken to gather information on health indicators. It aims to provide reliable estimates for fertility, health indicators and infant mortality (Haslett et al., 2014b). It is collected every five years, with the first one being conducted in 1996. There were 10,888 households surveyed, with a 99% response rate, but anthropometric data was only collected for a sub-sample of households, resulting in 2,345 children under the age of five being included in the survey.

The design was a two stage stratified cluster sample; the strata were the urban and rural parts of the 13 geographical domains, and the PSUs were wards. Wards were selected within strata using PPS. The selection ratio of rural to urban PSUs was roughly 2:1. Thirty-five households were sampled in each urban PSU and 40 households were sampled in each of the rural EAs. For further details on the DHS and the sampling method see Ministry of Health and Population (MOHP)[Nepal] (2012) and Haslett et al. (2014b).

Table 4.4: Structure of the NDHS2011.

| | Region | District | Ilaka | VDC/Mun | Ward |
|---|---|---|---|---|---|
| Contains | 5 | 72 | 215 | 233 | 283 |
| Mean Children | 478 | 33 | 11 | 10 | 8.5 |
| Min Children | 346 | 3 | 1 | 1 | 1 |
| Mean Household | 362 | 25 | 8.4 | 7.8 | 6.4 |
| Min Household | 275 | 3 | 1 | 1 | 1 |
| Mean ea | 58 | 4 | 1.3 | 1.2 | 1.02 |
| Min ea | 43 | 1 | 1 | 1 | 1 |

### 4.2.3 Model Formulation

This is an outline of the method used by Haslett et al. (2014b) to form the model to predict small area estimates of wasting rates. The DHS data was used to fit a model to the standardized

wasting rate of children under five (WHZ). There were 66 potential variables that could be fitted in the model as well as 12 GIS variables. The variables included a combination of numerical predictors and categorical predictors, and the categorical predictors were made into binary variables. Taking into account possible variable interactions, there were a large number of possible models that could be fitted; consequently the model fitting was done initially via an automated process. In general, hierarchical modelling was employed: an interaction between two variables was only included if the main effect for each of the variables was included. Regional interactions were included when a variable had different effects between the regions. It was important that the number of variables fitted to the model was kept relatively small, otherwise there may have been a risk of overfitting.

The final model developed from the initial model fits had 22 parameters with no interactions between variables being included. The variables from the selected model are defined in Table 4.7 and the model is shown in Table 4.8. Of these variables, 12 of them are ward level means, where this is denoted by a "W" in the variable names, (wards are subclusters in the DHS). For example $Wwater\_piped$ is the proportion of households in the ward that have their water pumped. The coefficient of determination was very low where the model explained only 11.5% of the variation in the data. However, most of the unexplained variation is between children within a household and not between the clusters; the between cluster coefficient of determination was 74.03%.

The final WHZ were produced following the full ELL process described in Section 3.3.1, with the exception that heteroscedasticity was not adjusted for in the household and child level residual variance as it was not found significant. The full process and final small area estimates are outlined in Haslett et al. (2014b).

## 4.3 Appendix

Table 4.5: Variable definitions in the Cambodian poverty model.

| Variable | Variable Description |
|---|---|
| hhsize | household size |
| lnhhsz | natural log of household size |
| pkids06 | prop of hh aged 0-6 |
| plit | prop of hh literate |
| pseced | prop of hh with secondary education |
| notoilet | no toilet within premises |
| numroom | number of rooms |
| rfree | dwelling is rent free |
| car | number of cars owned |
| cellphone | number of cellphones owned |
| computer | number of computers owned |
| electric | main source of lighting is electricity |
| motorbike | number of motorbikes owned |
| phone | number of phones owned |
| radio | number of radios owned |
| tv | number of tvs owned |
| floor_t | floor of tiles |
| floor_c | floor of cement,parquet |
| floor_s | floor of stone |
| roof_t | roof of tiles |
| roof_c | roof of concrete,other |
| roof_m | roof of metal |
| wall_b | walls of bamboo/mixed type |
| boat_e | mean number of boats owned |
| cellphone_e | mean number of cellphones owned |
| h_lit_e | propn hhead literate |
| plit_e | prop of ea literate |
| resplus_e | propn hh residential+shop/business |
| reg3 | rural (outside Phnom Penh) |
| tonlesap | Tonlesap ecological zone |
| plnmount | Plains/Mountains ecological zone |
| hhsizeXS3 | interaction of hhsize and rural |
| roof_cXS3 | interaction of roof_c and rural |
| numroomXS3 | interaction of numroom and rural |
| motorbikeXS3 | interaction of motorbike and rural |
| _cons | constant term |

Table 4.6: Fitted regression model for Cambodia.

| ln_exp | Coef. | Std. Err. | t | P>t |
|---|---|---|---|---|
| hhsize | -0.0344 | 0.0086 | -4.01 | 0.000 |
| lnhhsz | -0.5469 | 0.0339 | -16.11 | 0.000 |
| pkids06 | -0.1092 | 0.0254 | -4.31 | 0.000 |
| plit | 0.1141 | 0.0189 | 6.05 | 0.000 |
| pseced | 0.0843 | 0.0196 | 4.30 | 0.000 |
| notoilet | -0.0499 | 0.0144 | -3.46 | 0.001 |
| numroom | 0.0959 | 0.0123 | 7.81 | 0.000 |
| rfree | -0.1203 | 0.0242 | -4.97 | 0.000 |
| car | 0.2653 | 0.0252 | 10.54 | 0.000 |
| cellphone | 0.1230 | 0.0073 | 16.96 | 0.000 |
| computer | 0.1007 | 0.0255 | 3.95 | 0.000 |
| electric | 0.0512 | 0.0216 | 2.37 | 0.018 |
| motorbike | 0.0850 | 0.0135 | 6.27 | 0.000 |
| phone | 0.1254 | 0.0542 | 2.31 | 0.021 |
| radio | 0.0191 | 0.0087 | 2.19 | 0.029 |
| tv | 0.0745 | 0.0095 | 7.87 | 0.000 |
| floor_t | 0.1029 | 0.0226 | 4.55 | 0.000 |
| floor_c | 0.0237 | 0.0063 | 3.76 | 0.000 |
| floor_s | 0.4183 | 0.1662 | 2.52 | 0.012 |
| roof_t | 0.1122 | 0.0167 | 6.70 | 0.000 |
| roof_c | 0.0602 | 0.0322 | 1.87 | 0.062 |
| roof_m | 0.0558 | 0.0143 | 3.90 | 0.000 |
| wall_b | -0.0599 | 0.0123 | -4.86 | 0.000 |
| boat_e | 0.1522 | 0.0406 | 3.74 | 0.000 |
| cellphone_e | 0.1645 | 0.0243 | 6.77 | 0.000 |
| h_lit_e | 0.3222 | 0.1158 | 2.78 | 0.006 |
| plit_e | -0.4534 | 0.1367 | -3.32 | 0.001 |
| resplus_e | 0.2330 | 0.0786 | 2.96 | 0.003 |
| reg3 | -0.1529 | 0.0434 | -3.52 | 0.000 |
| tonlesap | -0.0597 | 0.0169 | -3.53 | 0.000 |
| plnmount | -0.0673 | 0.0250 | -2.69 | 0.007 |
| hhsizeXS3 | 0.0317 | 0.0057 | 5.56 | 0.000 |
| roof_cXS3 | 0.1587 | 0.0500 | 3.17 | 0.002 |
| numroomXS3 | -0.0289 | 0.0157 | -1.84 | 0.067 |
| motorbikeXS3 | 0.0359 | 0.0156 | 2.30 | 0.022 |
| _cons | 9.3097 | 0.0660 | 141.10 | 0.000 |

Table 4.7: Variable Definitions in the Nepalese wasting rate model.

| Variable | Variable Description |
|---|---|
| ageyr23 | age in years = 1 |
| girl | 1 if child is a girl |
| terai | 1 if HH is located in terai |
| wat_cwell | 1 if drinking water is from a covered well |
| hage2 | 1 if HH head aged 30-44 |
| flr_con | 1 if floor material is concrete |
| wall_wood | 1 if wall material is wood/planks |
| wall_bambo | 1 if wall material is bamboo |
| wall_brk | 1 if wall material is baked bricks |
| Wroof_iron | % HH with iron roof material, ward |
| Wroof_tile | % HH with tile roof material, ward |
| Wmax_educ_none | %HH with no educational attainment, ward |
| Whead_female | % of female headed HH, ward |
| Wroof_straw | % HH with straw roof material |
| Wmax_educ_fem_5to7 | %HH with maximum female educational attainment 5-7 years, ward |
| Wtoilet_flushseptik | %HH with flush to septik toilets, ward |
| Wroof_mud | %HH with mud roof material, ward |
| Wtoilet_none | %HH with no toilet, ward |
| Wwater_piped | %HH with piped water, ward |
| Wowns_fridge | %HH with fridge, ward |
| meanht | mean ht of VDC |
| popdens | population density of VDC |

Table 4.8: Fitted regression model for the Nepalese wasting rate.

| WHZ | Coef. | Std. Err. | t | P>t |
|---|---|---|---|---|
| ageyr23 | -0.1285 | 0.0563 | -2.28 | 0.023 |
| girl | 0.1085 | 0.0473 | 2.29 | 0.023 |
| terai | 0.4378 | 0.0824 | 5.31 | 0.000 |
| wat_cwell | 0.3943 | 0.1734 | 2.27 | 0.024 |
| hage2 | -0.1533 | 0.0580 | -2.64 | 0.009 |
| flr_con | 0.3781 | 0.0997 | 3.79 | 0.000 |
| wall_wood | 1.3413 | 0.3403 | 3.94 | 0.000 |
| wall_bambo | 1.2164 | 0.3157 | 3.85 | 0.000 |
| wall_brk | 1.2109 | 0.3139 | 3.86 | 0.000 |
| Wroof_iron | 1.0195 | 0.1901 | 5.36 | 0.000 |
| Wroof_tile | 1.0805 | 0.2030 | 5.32 | 0.000 |
| Wmax_educ_none | 0.8996 | 0.2166 | 4.15 | 0.000 |
| Whead_female | 0.5263 | 0.2250 | 2.34 | 0.020 |
| Wroof_straw | 1.0849 | 0.2222 | 4.88 | 0.000 |
| Wmax_educ_fem_5to7 | 2.4922 | 0.6705 | 3.72 | 0.000 |
| Wtoilet_flushseptik | -0.3756 | 0.1300 | -2.89 | 0.004 |
| Wroof_mud | 0.7599 | 0.2435 | 3.12 | 0.002 |
| Wtoilet_none | -0.7830 | 0.1229 | -6.37 | 0.000 |
| Wwater_piped | 0.2196 | 0.0892 | 2.46 | 0.014 |
| Wowns_fridge | 1.9809 | 0.5794 | 3.42 | 0.001 |
| meanht | 0.1940 | 0.0612 | 3.17 | 0.002 |
| popdens | 0.0000 | 0.0000 | 2.59 | 0.010 |
| _cons | -3.4999 | 0.4725 | -7.41 | 0.000 |

# Chapter 5

# A Variable Importance Metric for Small Area Estimates

This chapter reviews the concept of variable importance and proposes a method to measure variable importance specifically for small area estimation. Variable importance helps to determine the ranking of variables in terms of their effect on a fitted model. It measures the practical significance a variable has in a fitted model. Small area models can often be very complex. However a large number of variables may be of little use if they provide no predictive power at small area level. If variables are not important in the final small area estimate they may not need to be included in the model. This adds another criterion to model selection, which is not usually considered. Most methods focus on how important a variable is in fitting a regression model to the response, however I focus here on a variable's importance in predicting the small area estimate. This leads to a method for reducing model complexity in SAE without significantly changing estimated levels or estimated mean squared error for each small area. Assessments of a variable's importance developed here consider not only each auxiliary variable's ability to explain unit-level variation in the dependent variable (usually assessed via F-tests), but also its

ability to distinguish between relative levels in the small areas and the effect of its deletion on SAE accuracy. The diagnostic developed covers a wide range of SAE methods, including those based on survey data only and those which combine survey and census data. The core question addressed is how candidate survey-based models might be simplified without losing accuracy or introducing bias into SAEs. Using a novel ordering of the effects in the model, based on their direct influence on the small area estimates, I illustrate how to assess simplification of SAE models, while avoiding marked changes in the estimated level for each small area or loss of precision. The diagnostic method is illustrated using estimation of commune-level poverty rates in Cambodia from national household-level data.

## 5.1 Introduction

Models are only useful if they provide predictive power for the response variable. Although a model may be complex, it has little value if the explanatory variables are not able to predict variations in the quantity of interest. In this case a metric to evaluate the importance of a variable becomes useful in order to determine the relative importance of each of the regressor variables and to aid in assessing whether they are adding anything to the predictive power of the model. In linear models some of the common measures used in model selection include residual mean squared error, Mallows Cp, $R^2$ adjusted, AIC and BIC. These measures all take into consideration either the sum of squared residual error or the residual mean square error from the fitted model, as well as penalising for the number of parameters in the model. Therefore, if a variable is not helping to explain a significant amount of the variation in the response variable after taking into account other variables then this will be reflected in the change in the value of the various measures when the variables are added/deleted. Grömping (2015) outlines a wealth of references related to variable importance in linear models as well as outlining various

variable importance metrics including analysing the methods for variance decomposition, non-linear models and methods for machine learning. Grömping (2015) also gives an overview of the diversity of concepts for variable importance that is based on the work of Achen (1982), who outlines importance measures for linear unit level models. These include level importance; this combines the unstandardised estimated regression parameter $b_p$, which is the influence on the response from a one unit change in the $p^{th}$ variable, and the average across the $p^{th}$ variable $(X_p)$. This results in a measure of $b_p.\bar{X}_p$. The other measure is the dispersion measure, where this standardizes the estimated regression coefficient. However, these metrics are not always practical in SAE as small area models tend to be generated from various types of mixed models. Vaida and Blanchard (2005) proposed conditional AIC, which is adapted for cluster effects and therefore for selection in mixed effect models. It does this by giving a penalty term that is related to the number of degrees of freedom in the linear mixed model. Van den Brakel and Buelens (2014) look at covariate selection for SAE in repeated sample surveys, as often different surveys will result in different model selection. They propose a methodology that performs model selection for all the survey editions simultaneously and selects the model by minimising the average AIC for all survey editions. Lahiri and Suntornchost (2015) focuses on variable selection in linear mixed models with applications to SAE. Often approximation error is evident in small area estimation as the real data is not collected but rather proxy data that includes error. Approximation error occurs when the difference between the standard variable selection and the variable selection that would occur in the presence of no sampling error does not converge to zero. When this occurs, they propose an adjustment to the Fay-Herriot method that reduced the approximation error. Diagnostics in general tend to focus on the data used to fit the model, whereas the focus in this application is on diagnostics for the final small area estimate, which typically involves auxiliary data in addition to the data used in the model.

Current variable importance methods assume that we are only interested in the effect the

auxiliary variable has on the unit-level response variable. In SAE this is not the case, as the response variable at unit level is then aggregated up to small area level. A variable may appear to be beneficial at the unit level, but if it lacks diversity or variation between the small areas it will not be adding anything to the overall small area estimates. The importance of a variable in small area estimation is dependent not only on its predictive power at the unit level but also its ability to differentiate between the small areas. Note that in a linear model a variable that had the same mean for every area $i$ would not contribute to the variability between estimates even if it was highly significant in the regression model. For example, in the developing world, when modelling the proportion of children in each small area considered to be underweight, via standardized weight for age, age is often an important variable in determining whether an individual child is underweight and shown as highly significant in the regression model. However, when the child level predictions are aggregated to small area level the variable is of limited value for distinguishing between the small areas, as there is typically the same proportion of children of each age in each small area.

In this chapter, a metric for variable importance is proposed that looks at the contribution of not only the standardized coefficient of the regression model but also the variability of the mean of the explanatory variable between the small areas. This is somewhat similar to the metric proposed by Achen (1982) used for linear unit level models ($b_p.\bar{X}_p$). However the proposed metric in this chapter combines the regression coefficient with the dispersion of the mean of the variable between the small areas. Small area estimation differs from unit linear regression as the unit level predictions need to be amalgamated up to small area level.

When there is unit level census data from a population of size $N$ available, then the resulting matrix of covariates $\mathbf{X} = [x_{ij}]_{N\times P}$ and the unit-level predictions ($\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$) need to be amalgamated to small area level. This amalgamation can be represented in matrix form as

$\hat{\bar{Y}} = \mathbf{A}\hat{\mathbf{Y}}$ where $\hat{\bar{\mathbf{Y}}} = (\hat{\bar{y}}_1, ..., \hat{\bar{y}}_p, ..., \hat{\bar{y}}_I)'$ is the $I \times 1$ vector of small area estimates, and

$$\mathbf{A} = \begin{bmatrix} \frac{1}{N_1} & & & & \\ & \ddots & & & \\ & & \frac{1}{N_i} & & \\ & & & \ddots & \\ & & & & \frac{1}{N_I} \end{bmatrix} \begin{bmatrix} \mathbf{1}_1' & \cdots & \mathbf{0}_i' & \cdots & \mathbf{0}_I' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_1' & \cdots & \mathbf{1}_i' & \cdots & \mathbf{0}_I' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_1' & \cdots & \mathbf{0}_i' & \cdots & \mathbf{1}_I' \end{bmatrix} = \begin{bmatrix} \frac{1}{N_1} & & & & \\ & \ddots & & & \\ & & \frac{1}{N_i} & & \\ & & & \ddots & \\ & & & & \frac{1}{N_I} \end{bmatrix} \begin{bmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_i' \\ \vdots \\ \mathbf{a}_I' \end{bmatrix} \quad (5.1)$$

where $\mathbf{1}_i$ and $\mathbf{0}_i$ are subvectors of ones and zeros respectively of length $N_i$, $\mathbf{a}_i = (\mathbf{0}_1', \mathbf{0}_2', ..., \mathbf{0}_{i-1}', \mathbf{1}_i' ... \mathbf{0}_I')'$, and where $N_i$ is the population size in the $i^{th}$ small area.

Hence $\hat{\bar{\mathbf{Y}}} = \mathbf{A}\mathbf{X}\hat{\beta} = \bar{\mathbf{X}}\hat{\beta}$ where

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{x}_{11} & \cdots & \bar{x}_{1p} & \cdots & \bar{x}_{1P} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{I1} & \cdots & \bar{x}_{Ip} & \cdots & \bar{x}_{IP} \end{bmatrix} \quad (5.2)$$

is the $I \times P$ matrix of area-level covariate means.

When considering the variability of the small area estimates in $\hat{\bar{Y}}$, the focus needs to be on the variability of the auxiliary variables between the small areas for a given set of estimated parameter $\hat{\beta}$. The relevant quantity can be expressed as:

$$V(\bar{\mathbf{X}}\hat{\beta}|\hat{\beta}) \quad (5.3)$$

This variance is with respect to its definition in survey methodology where it measures the spread of a set of fixed numbers from the average, rather than referring to the definition of variance with respect to a stochastic process. Because the interest is in predictor variables that vary across small areas, the focus is on the variance of the elements in each column of $\bar{\mathbf{X}}$, i.e.

for each variable in the regression, where

$$V(\bar{x}_{ip}) = \frac{1}{I-1}\sum_{i=1}^{I}(\bar{x}_{ip} - \bar{x}_{.p})^2 \tag{5.4}$$

and $\bar{x}_{.p}$ is the overall population mean for variable $p$; here '.' is used to indicate the arithmetic mean over an index. Equation (5.4) is relevant for both area level and unit level models. When using area level data $\bar{x}_{ip}$ are the area level covariates and when using unit level data, it is the aggregation of unit data to small area level. The contribution of variable $p$ to the overall variability in (5.3) is then

$$V(\bar{x}_{ip}\hat{\beta}_p|\hat{\beta}_p) = V_p(\bar{x}_{ip})\hat{\beta}_p^2 \qquad p = 1,2,...,P \tag{5.5}$$

The variability between the small areas in (5.5) measures the usefulness of each variable in terms of how well it distinguishes between the small areas in a given fitted regression model. This is further expanded in Section 5.2.1. This differs from the variance of the regression parameters; $V(\hat{\beta})$; which allows assessment of $\hat{\beta}$ via F-tests and measures the uncertainty in the regression parameter estimates from the survey data. While important in model selection, this is not sufficient for SAE because $\hat{\beta}_p$ is unchanged from small area to small area, and the focus of a variable in this application is also to assess the variability between small areas. The variance (5.5) plays a crucial role in assessing the variables that are important in small area estimation, as opposed to their importance in the regression model fitted to the survey data.

Although the proposed diagnostic has been developed for the ELL method, it is able to be applied to other small area estimation methods such as the EBLUP where the estimate is made up of a direct estimate and a synthetic estimate. Referring back to the EBLUP equation in (2.5), this can be rewritten as

$$\hat{\theta}_i = \bar{x}_i'\hat{\beta} + \hat{u} \tag{5.6}$$

where $\hat{u}$ is an estimated area-level effect. Therefore the EBLUP and other SAEs can be regarded as an adjustment of the synthetic estimate $(xbar_i' betahat)$, using an estimated area-level effect. This could also be expressed as:

$$\hat{\theta}_i = \frac{\sigma_e^2}{\sigma_e^2 + n_i\sigma_v^2}\vec{x}_i'\hat{\beta} + \frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}\bar{y}_i \tag{5.7}$$

where this a weighted average of the synthetic estimator $\vec{x}_i'\hat{\beta}$ and the direct estimate $\bar{y}_i$. If $n_i$ is small then the direct component of the estimate is small and the synthetic component of the estimate is weighted higher. If the sample size is the small areas is small and so the synthetic component has a large weight and the diagnostic would behave similarly to those applied to estimates that were entirely synthetic, such as the ELL. In situations where there is a larger amount of sampled data in the small area, the diagnostic may behave in a different manner. The diagnostics described in the remainder of this thesis can be applied to SAE methods such as the EBLUP where synthetic and direct estimates are combined, but will just apply to synthetic portion of the estimate. The performance when the sampled size is larger may be a further area of research.

## 5.2 Methodology

### 5.2.1 Ranking of Contextual or Auxiliary Variables in SAE

Using the measure of variability between the small areas explained above in (5.5) a new method for ranking the auxiliary variables in terms of their importance in SAE is proposed. Instead of using only the $t$-statistics and $F$-statistics to select the model fitted to the survey data, the variation in the explanatory variables between the small areas is also taken into account. As an initial step, a new variable $(\hat{\tau}_{ip})$ is defined. This combines the mean of the $p^{th}$ regressor in a

small area $i$ as well as the standardised regression coefficient for that variable $\hat{\beta}_p$; where this is fitted using the survey data. This is shown as:

$$\hat{\tau}_{ip} = \bar{x}_{ip}\hat{\beta}_p. \tag{5.8}$$

When census data is available for all the auxiliary variables, there is an estimated $\hat{\tau}_{ip}$ for each variable in each small area, resulting in $I \times P$ values; where $P$ is the number of variables and $I$ is the number of small areas. For this situation, there is no uncertainty in $\bar{x}_{ip}$ as there is information for each member of the population from a census. The situation where only survey and no census data is available, and each $\bar{x}_{ip}$ is to be estimated, will be illustrated later in the chapter.

There are two components of variability in $\hat{\tau}_{ip}$; one is the conditional variance given the small area $i$ and the other is the conditional variance given the regression coefficient $\hat{\beta}_p$. These are defined as:

$$V(\hat{\tau}_{ip}|\bar{x}_{ip}) = \bar{x}_{ip}^2 V(\hat{\beta}_p) \tag{5.9}$$

and

$$V(\hat{\tau}_{ip}|\hat{\beta}_p) = V(\bar{x}_{ip})\hat{\beta}_p^2 \tag{5.10}$$

respectively. The first is not of immediate focus because although it is important for finding atypical small areas, $V(\hat{\beta})$ remains fixed across the small areas and focuses on measures of uncertainty in the regression parameters. The second source of variability shown in (5.10) is a restatement of (5.5). This measures the $p^{th}$ auxiliary variable's contribution to the small area estimates conditional on the fitted model.

In the unit level model, a variable is usually considered significant and is included in the model if it has a t-statistic greater than the critical value; or similar methods such as assessing

the contribution to $R^2$ adjusted. For small area estimation, I instead propose the importance of a variable be determined not only by the significance in its t or F-statistic but also by (5.10), where candidate variables with larger contributions to between area variance are given a higher importance. By ranking the variables based on this criterion, it will assist in determining which variables should be removed first if model simplification is desired. Note that in (5.10), auxiliary variables in the model do not need to be standardized as the product is invariant to scale.

Because the variance is additive, whereas the standard deviation is not, the variance of each variable $p$ across the small areas is initially considered. Later the square root of the variance is taken to give $SD(\hat{\tau}_{ip}|\hat{\beta}_p)$, which is defined to be the variable importance measure (VIM).

### 5.2.2 Only Survey Data Available

There are many situations in SAE when unit record census data is unavailable, in which case the variability between the small areas can be estimated using survey data. The survey may contain samples from all the small areas or just a selection of small areas. In either situation (5.5) cannot be used, as neither $\bar{x}_{ip}$ for $i = 1, ..., I_s$ nor $\bar{x}_{.p}$ are known and can only be estimated. A naïve estimate of the variability across the small areas for variable $p$ is

$$\hat{V}(\hat{\bar{x}}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}\sum_{i \in s}^{I_s}(\hat{\bar{x}}_{ip} - \hat{\bar{x}}_{.p})^2 \tag{5.11}$$

where there are $I_s$ small areas sampled in the survey, $\hat{\bar{x}}_{ip}$ is the survey weighted sample mean of small area $i$ for variable $p$; and $\hat{\bar{x}}_{.p}$ is the overall weighted sample mean for a particular variable. The complication is that using only the survey data in (5.11) makes this a biased estimate for the variance (5.5) even if all small areas are sampled. The variance is overestimated because (5.11) includes not only the between small area variability but also the within small area variability and the uncertainty of the true mean from the population and the estimated mean

from the sample (for more details see the Appendix A). An unbiased estimate of the variance requires knowledge of the design weights at each sampling level (Särndal et al., 1992, p.137) or knowledge of the joint inclusion probabilities as well as the population size in each small area (Korn and Graubard, 2003). However, usually only the final level survey weights are available, in which case only an approximate decomposition of the variance into the between and within components is possible. One possibility is the decomposition method outlined in the appendix of Elbers et al. (2002), which provides a practical approximate correction for the within-area variation.

The estimated sample variance generated in (5.11) will be approximately unbiased and the ranking of the auxiliary variables will remain the same if the variability between the small areas is low. The variability for SAE models based only on survey data is investigated using both the naïve estimate (5.11) and the Elbers et al. (2002) approximation method in Section 5.3.5.

## 5.3 Application to Cambodia Data

The variable importance metric is applied to the Cambodian dataset outlined in Chapter 4. As a starting point, I used the final fitted model by Haslett et al. (2013). This model was used to predict the log expenditure of a household with these predictions later used to predict the small area poverty rates. This initial model has 35 variables, which were all deemed significant in terms of their ability to explain an individual's log expenditure. These variables and their definitions are outlined in Table 4.5, the fitted regression model is shown in Table 4.6 and the correlation matrix for the covariates and the response in Appendix 5.2 in Tables 5.2-5.5. Aggregation of the linear model gives the mean of log expenditure, however poverty is a non-linear function of the response $log(expenditure)$, so it needs to be investigated whether the

diagnostic still works in the case of the non-linear function.

In this investigation, the variable with the lowest value of $sd(\hat{\tau}_{ip}|\hat{\beta}_p)$ was removed and the model refitted. The process continued each time the model was refitted. The values of $sd(\hat{\tau}_{ip}|\hat{\beta}_p)$ slightly change, as some variables became more important and others less important due to the interdependent relationship between the variables. Table 5.1 in the Appendix shows the VIM, $sd(\hat{\tau}_{ip}|\hat{\beta}_p)$, of each variable in the fifth column. The table shows the order the variables were removed from the model. The VIM cannot be negative, as it is impossible to have a negative standard deviation. The first five variables removed remain in the same order as if the model was not refitted, as even when the previous variables are removed they are still less important that the others. In both the original model and when the model is continually refitted, $cellphone\_e$ remains the most important variable. In general, most of the other variables only change a few positions in terms of the variable importance when the model is refitted compared to the original order. One of the variables that becomes more important is *pseced*, where this changes in place by six spots, as it becomes more important when other variables are removed. This process continued after the removal of each variable and the importance of each of the variables was reassessed until there was only one remaining variable. In Cambodia, this resulted in a total of 35 models being assessed. The small area estimates from this original model are considered to be the 'gold standard' and all the estimates from the reduced models were compared to these estimates. In an ideal world, one would know the true small area statistics and parameter values, however this is unrealistic so we assume that these original model estimates are the 'truth' or the closest to the truth possible.

In order to measure how the model and poverty estimates are affected from the model reduction technique several different diagnostics techniques are used. These consist of only looking at how the fitted linear model changes, followed by examining how the reduced model affects the corresponding poverty estimates (where these are a non-linear function of the re-

sponse) and the uncertainty surrounding these.

### 5.3.1 Initial Model Diagnostics

An initial step was used to assess the series of models using diagnostics that only consider the deterioration of the fit of the model and the magnitude of the unexplained variation based on the survey data. Although it does not formally make allowance for the model containing multiple random effects, $R^2$ gives an indication of the percentage of the variability in the dependent variable that is explained by the regression parameters based on the survey data. Figure 5.1 shows the deterioration in the first stage model fit in terms of $R^2$ as variables are sequentially removed from the model. This shows the decreasing proportion of variation in household level log expenditure explained by the model as variables are removed. Although $R^2$ is not a sufficient criterion, it is noteworthy that nine variables can be excluded from the model with the $R^2$ only decreasing from 0.655 to 0.650, as these nine parameters only explain a small percentage in the total variation. Fifteen variables can be removed from the initial model and the corresponding $R^2$ decreases by less than 0.01. Furthermore, when 24 variables are removed the fit of the model only decreases by 0.05. The largest decrease in $R^2$ occurs when the variable *lnhhsize* is removed from the regression, where the $R^2$ decreases from 0.5473 to 0.4316. This shows when the model is small in terms of the number of variables it contains, *lnhhsize* is an important variable to include, as the removal of this reduces the quality of fit of the first stage model quite substantially. One thing to note is that as more variables are removed, the remaining variables in the model can became more important due to the interdependencies between the variables.

After sequentially removing variables, the last three variables remaining are *lnhhisze*, *cellphone* and *cellphone_e*. These three variables all have relatively large variation between the small areas. The model fit decreases only slightly when *tv* is dropped from the model.

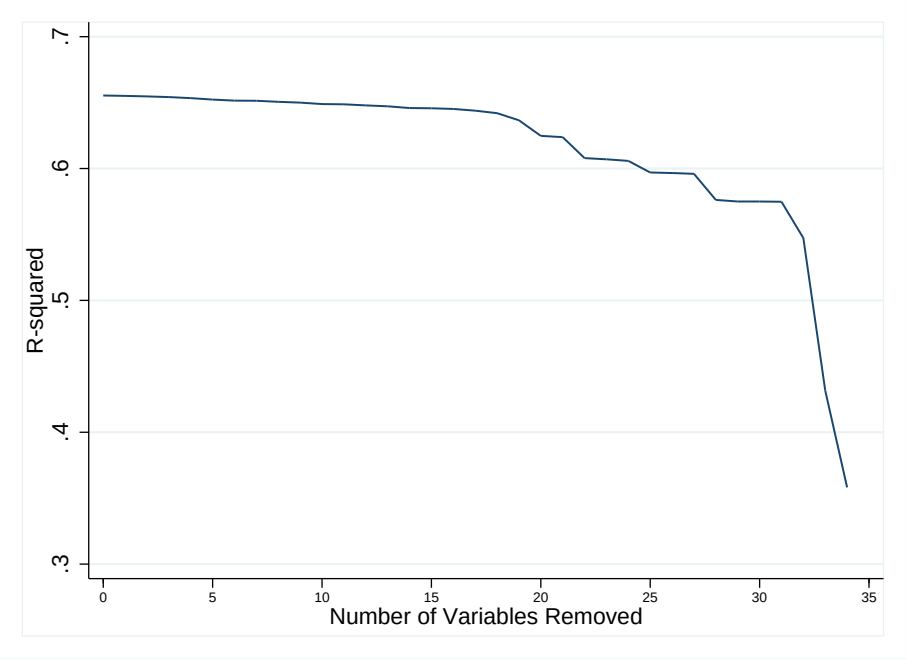



Figure 5.1: $R^2$ for the reduced survey based regression models.

However, there is a large decrease in the proportion of the response explained from 0.4316 to 0.3582 when *cellphone* is removed. The variable *cellphone_e* explains 35.8% of the variation in log expenditure at the household level; this contextual effect is an enumeration area (EA) level mean; which means it was collected at the cluster level. This means the proportion of people in the EA who own a cellphone helps to explain just over a third of the variability in log expenditure.

After examining the $R^2$, I also considered how both the cluster level variability and the ratio of the cluster level variation to the total variation deteriorated as the model was simplified. The examination of this is important because a useful strategy of small area estimation is reducing the ratio of the small area or cluster level error to the total error as well as minimizing cluster level variability. Reducing the proportion of the unexplained variability at the small area, or cluster level (especially when contextual variables are used) is beneficial, as unexplained random variation at the most aggregated levels typically has the biggest impact on

the precision of the small area estimates. Figure 5.2 illustrates that the ratio of the cluster level variance to total variance remains relatively constant when the first nine variables are removed. This slightly increases when the following four variables are removed, these being *resplus_e*, *numroom*, *roof_m* and *roof_c*. After 20 variables are removed it is relatively inconsistent, which means that for the removal of some variables the household uncertainty increases at a higher rate compared to the cluster uncertainty and other times the cluster uncertainty increases at a higher rate. Following the removal of the $27^{th}$ variable the ratio decreases, meaning that the removal of these last eight variables increases the household level uncertainty more than the cluster level uncertainty.

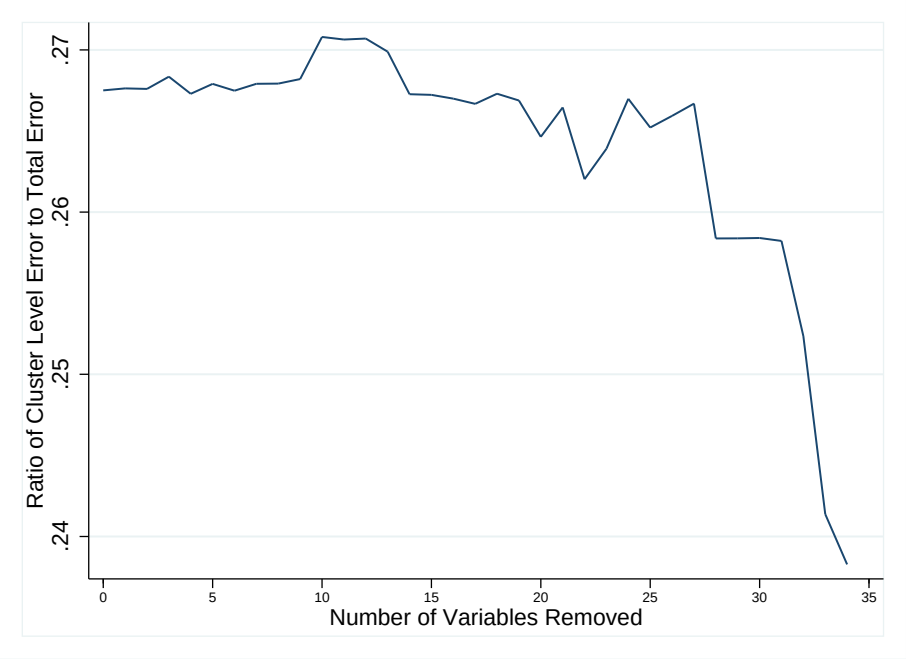



Figure 5.2: Ratio of unexplained cluster variance to total model error for the reduced models.

Figure 5.3 shows the change in the cluster level variance as the model is reduced. The figure shows that the unexplained cluster variance remains relatively constant at approximately 0.012 when the first nine variables are removed, with a slight increase when *resplus_e* is removed. It then remains relatively constant when the next eight variables are removed. After 19

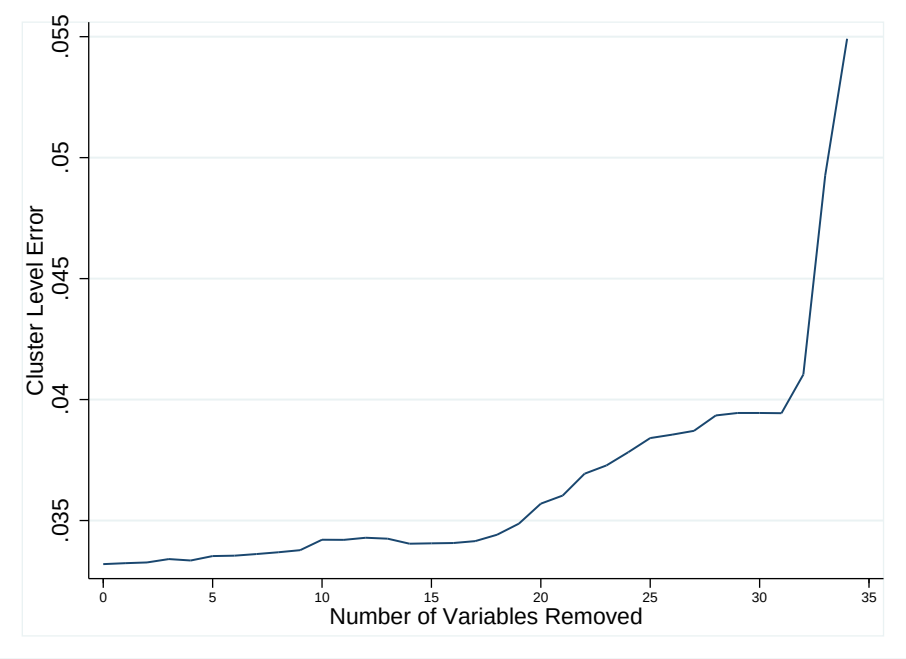



Figure 5.3: Unexplained cluster level variability for the reduced models.

variables are removed, the unexplained cluster level variability steadily increases with a rapid increase when *lnhhsize* and *cellphone* are removed, indicating that these models would be inappropriate for small area estimation, as there is considerable unexplained variation within the small areas.

The $R^2$ and the unexplained variability give an indication of the quality of the model for the survey data. However diagnostics to check the changes to the small area estimates are more important. These include quantifying how the poverty estimates in each of the small areas change with the reduced models as well as how the spatial distribution of poverty changes.

### 5.3.2 Point Estimates and Standard Errors

The initial model diagnostics were based on only refitting the model and observing the changes in the model and unexplained variability. In this section, diagnostics are based on running the

full ELL model process outlined in Section 3.3.1, where the model is fitted and the corresponding model is used to make predictions on the poverty status of each person in the population. These are then aggregated up to small area level, as well as the uncertainty surrounding those estimates. Here the focus shifts from unit level response (log expenditure) to the aggregated small area estimates $Pov_i$, where this is a non-linear function of the response.

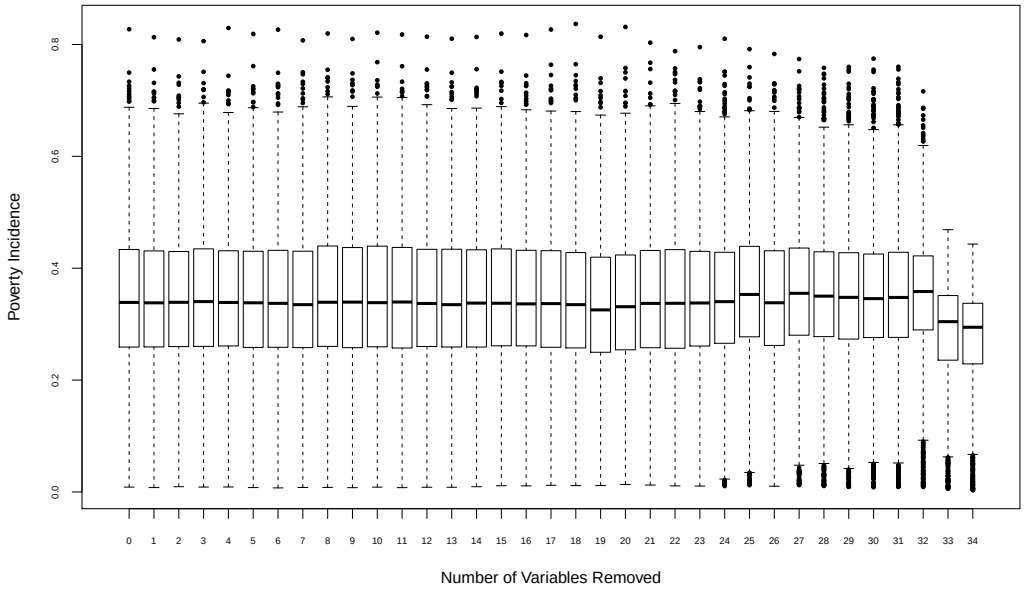



Figure 5.4: Point poverty estimates of the small area estimates, generated from the reduced models.

Figure 5.4 shows the distribution of the point estimates for the poverty incidence in each small area for a range of models. Each of the boxplots contains 1621 points, which represent the predicted poverty for each of the 1621 Cambodian communes. The distribution appears to be very similar for the majority of the models, with a slight change in distribution after the 23rd variable is removed. There is a noticeable change in the distribution of the SAE poverty point estimates after the 32nd variable is removed. In this case the point estimates of poverty tend to be lower than in the original model. Although this plot shows the distribution of the

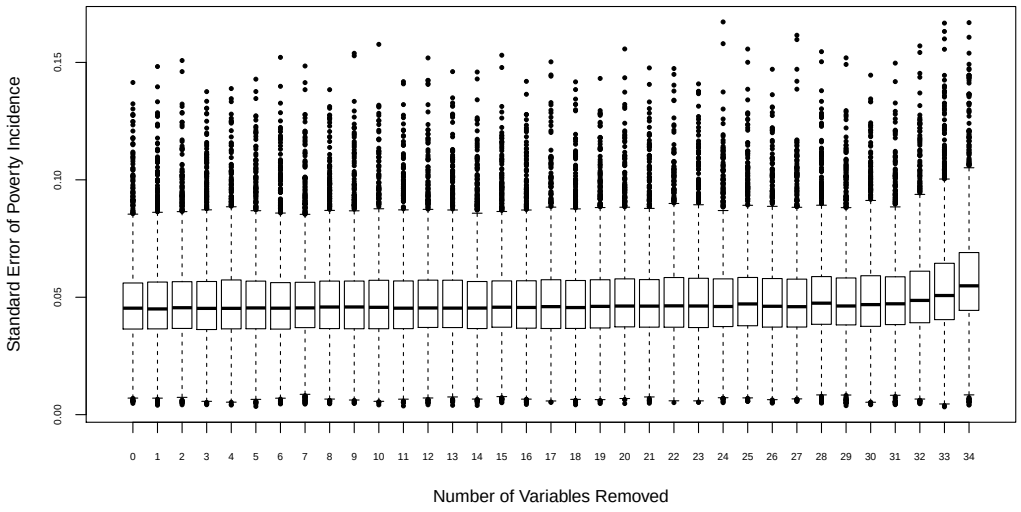



Figure 5.5: Standard error of the small area estimates, generated from the reduced models.

poverty estimates for all the communes (small areas) in Cambodia it does not allow us to make comparisons for particular small areas or understand how the poverty incidence has changed for a given commune. Even when the range of values is similar, the point estimate for each commune could be completely different. This would indicate that the poverty incidences for individual communes are sensitive to changes in the model. This is investigated in the next section.

The distribution of the estimated standard errors for the small areas can be seen in Figure 5.5. The spread of the standard errors for each model remains relatively constant. Estimated standard errors are conditional on the model being correct, and if the model is incorrect then the standard errors may be biased and not accurately reflect the model's uncertainty. We see here that the standard errors are robust to the choice of the model. If the error structure is misspecified then the precision of the estimates may be biased, i.e. the calculated standard errors would not be a true reflection of the uncertainty in the incidence of poverty

### 5.3.3 Comparison with Full Model Estimates

A comparison of the poverty predictions for each small area are compared for each of the simplified models with $r$ variables removed to the small area poverty predictions generated by the initial model; this being the model where no variables have been removed. A diagnostic was formed that took into account both the change in the poverty estimates as well as the uncertainty surrounding them. This was used to quantify the magnitude of each change in SAE as well as determining if the poverty estimates are significantly different. The diagnostic is shown in (5.12), where subscript $r$ indicates the number of variables removed in the reduced model

$$\zeta_r = \frac{Pov_0 - Pov_r}{\sqrt{se_0^2 + se_r^2}}. \tag{5.12}$$

Here $Pov_0$ is the poverty estimate with no variables removed from the initial model and $Pov_r$ is the poverty estimate with $r$ variables removed and $se_0$ and $se_r$ are the respected standard errors of the estimates. This diagnostic is a conservative approach because the real standard errors will be larger than those given by the denominator of (5.12), therefore it would result in a value of $\zeta$ larger than the true one, possibly leading to the conclusion that an unimportant variable is important and therefore not removed from the model; this is better than discarding a variable that may be important. Furthermore independence between $Pov_0$ and $Pov_r$ has been assumed in the conservative approach, however there is likely to be a positive correlation between the two estimates. Although correlation is present, it should be small as the main contributors to the standard error is the random effects, and the only correlation in the estimates is coming from the correlation in $\hat{\beta}$. For each model, there are 1621 values of $\zeta$ calculated, one for each small area. In (5.12) the subscript $i$ (denoting each small area) has been suppressed for clarity. This statistic allows one to observe if a commune's poverty estimate for a simplified model is significantly different to the full model.

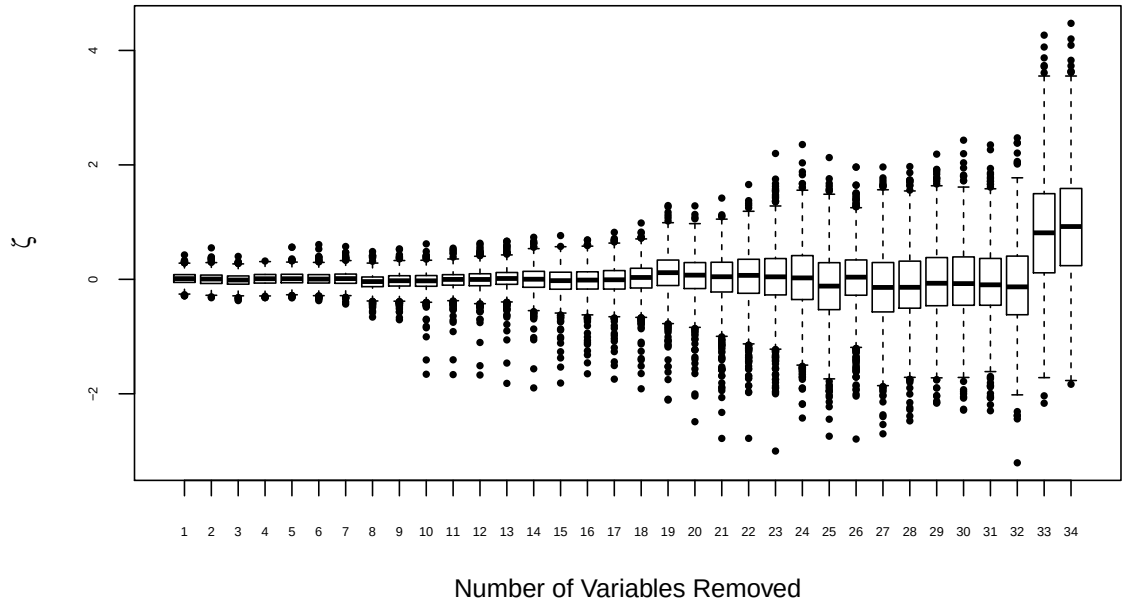



Figure 5.6: Difference in the small area poverty estimates between the original model and reduced models.

Figure 5.6 shows the distribution of the $\zeta-$statistic for each of the models with $r = 1$ to $r = 34$. It is important to note that the true small area statistics are unknown, rather we use assume that the initial fitted model estimates are a close representation of the true values, therefore the comparison is with the model estimates generated from the full model, rather than the true estimates. The figure shows there is only a small change in the predicted poverty incidence for each small area when the first seven variables are removed from the model. The value of $\zeta$ only changes by a maximum of 0.4 this would indicate that although these variables are explaining a significant amount of variation at the model fitting stage, they are not helping to distinguish between the predicted small areas. This could be a reflection of the variables having a similar value in all the same areas, or it contains a lot of variation within a particular small area. In both these situations, the variable(s) would not be helping to differentiate between

the small areas. When ten or more variables are removed, the estimated poverty rates begin to become considerably more inaccurate where we can see some small areas have a diagnostic statistic over 2. Although the middle 50% of the predictions are still very similar, there are several outlying points, suggesting it would be unwise to remove more than nine variables from the model. Removing more than ten variables would cause unreliable estimates for some of the small areas. When there are only one or two variables in the model all of the small areas start to become inaccurate on average, this can be shown by the median diagnostic shifting from approximately zero in the less reduced models to approximately 0.5. In general, as more variables are removed from the model the precision of the estimates begins to deteriorate; this is shown by the increasing spread of the $\zeta$. This suggests the more the model is reduced the less reliable the estimates are. This however is based on the assumption the original model estimates are true, when in reality the true values are unknown.

### 5.3.4 Correlation in the rank of the communes

In poverty mapping, one of the main focuses is on the spatial distribution of the poverty incidence within the country, as it is important that the areas of the country that are the poorest get targeted funding and aid. Therefore it is important that if the model is reduced the areas retain their relative order of poverty (e.g. the poorest small area in the original model is also the poorest small area in the simplified model). Although the point estimates of the poverty incidence may not be significantly different for two models, we cannot know from the point estimates alone if the relative order of the communes in terms of the severity of poverty is similar. To test the spatial distribution of poverty for the small areas, a non-parametric Spearman rank correlation test was used to compare the relative ranking of the communes in terms of the severity of poverty under the different models. The correlation is used to identify the relationship between the ordering of the communes in terms of their incidence of poverty.

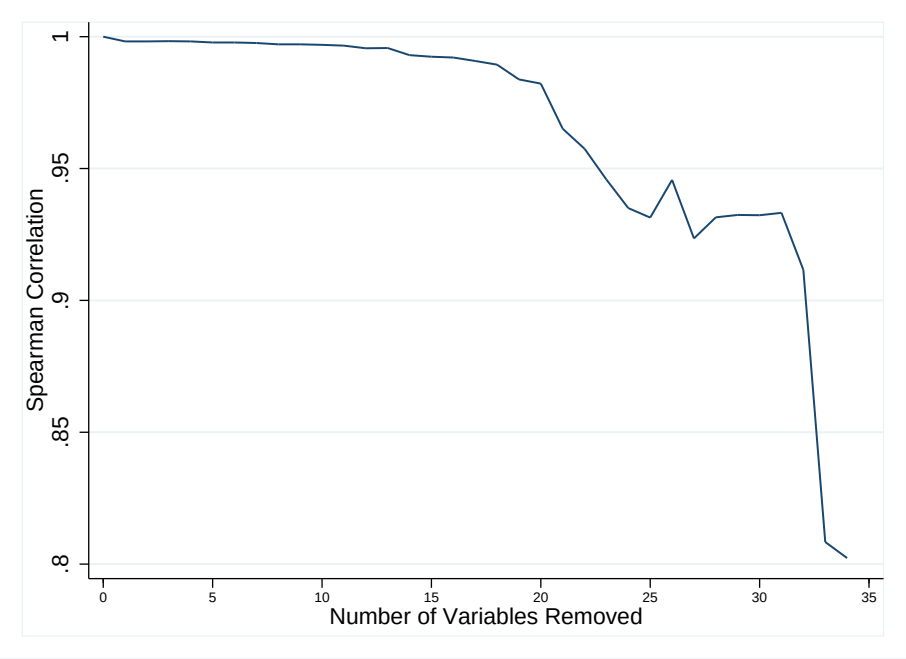



Figure 5.7: Spearman correlation rank for the small area poverty estimates in the original model compared to the reduced models.

Figure 5.7 illustrates that the ordering of the small areas is robust to reductions in model dimensionality; this can be seen as the Spearman rank correlation remains above 0.99 for the first 17 models, indicating 17 variables can be removed and the ordering of the communes in terms of their incidence of poverty remains relatively the same. Even after 31 variables are removed the rank of the communes in terms of their level of poverty are still relatively similar to the full model, with the Spearman rank correlation being 0.933.

There is a relatively large decrease in the Spearman rank correlation with the removal of the 33$^{rd}$ variable which is the log of the household size. This indicates that the log of the household size is useful as it not only explains a large amount of the variation at household level in the survey data but it also helps distinguish between the small area estimates.

If we are only concerned about the ordering of the communes in terms of the level of poverty, we could remove 17 variables and simplify the model to having only 18 variables. This

would result in the ordering remaining relatively constant. Up to 22 variables can be removed in order to have a Spearman rank correlation above 0.95. However, if we refer to Figure 5.6, caution is required as removing any more than 9 variables leads to changes in the estimated level of poverty for the communes, consequently removing too many variables will reduce the accuracy of the small area level poverty estimates.

### 5.3.5 Analysis using Survey Data Only

There are two main applications for considering survey data only:

- The first is when only survey data is available.
- The second is to provide a preliminary assessment of modelling. This reduces the extensive computation time required at the model checking stage.

As outlined in Section 5.2.2, it is difficult to get an unbiased estimate of the variance for $\hat{\tau}_{ip}$ across the small areas using the sample data alone. However having access to both Cambodia's survey and census data it is possible to examine the bias introduced by using only the survey data.

Figure 5.8 shows the comparison of $\{\hat{V}(\hat{\tau}_{ip}|\hat{\beta}_p)\}^{1/2}$ using the census and the sample data, versus the sample data only; this is based on the full model with 35 variables. If the sample generates unbiased variance estimates of the variability for each of the variables, the data points will fall along the 45 degree line. Any data points above the line indicate that sample variability is underestimated and any data points below the 45 degree line indicate that sample estimate of variance is over inflated.

The figure shows that for the majority of the variables the Elbers et al. (2002) approximation method produces estimates more accurate than using the naïve estimator in (5.11). This is

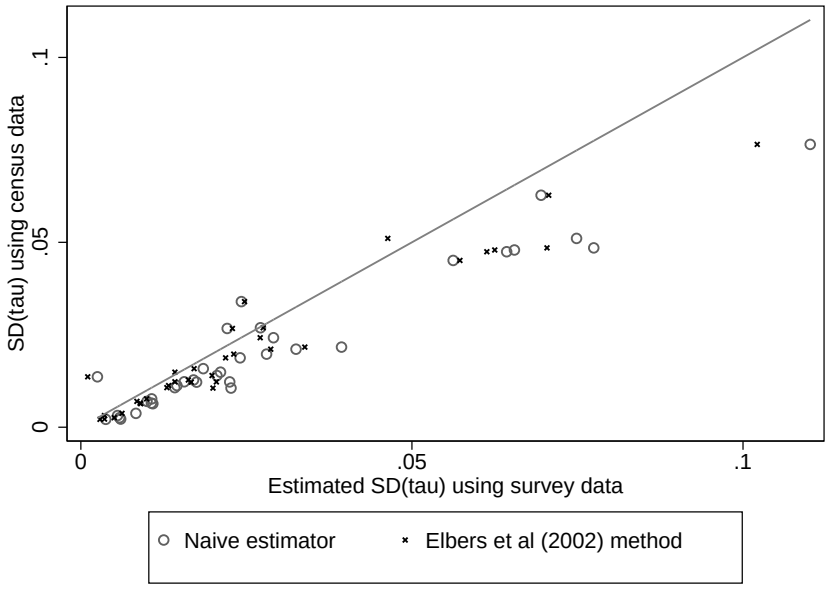



Figure 5.8: $SD(\hat{\tau}_{ip}|\hat{\beta}_p)$ of the survey using the naïve variance estimator (5.11) and Elbers et al. (2002) approximation.

shown by the points generated using the approximation method being closer to the 45 degree line. For the variables where there is not much variation within the small areas, the naïve and the approximation method are relatively similar. Because $\hat{\beta}_p$ is held fixed, the possible bias must come from estimating the variability in $\bar{x}_{ip}$ across the small areas. The primary reason the majority of the variance estimates being upwardly biased is due to there being a relatively large difference between $\hat{\bar{x}}_{ip}$ and $\bar{x}_{ip}$ in each small area (see Section 5.2.2). This is due to the sample size in each small area being relatively low, as seen by the sample size varying between 9 and 40. This small sample size does not accurately reflect the true value of the variable in each small area. At the survey design phase, one possibility would be to increase the sample size in each small area although this will reduce the number of small areas sampled. In the case of Cambodia, the communes are the desired small areas and there are only 621 of the 1621 communes with survey data available. It would be possible to amalgamate the communes and define the small areas as a higher geographical regions such as district, or province level, in which case the

larger sample size in each district and even larger in the province would help to reduce the bias in the estimate of $\hat{V}(\hat{\tau}_{ip}|\hat{\beta}_p)$. For Cambodia aggregating the data to province level would result in 75 small areas; this better reflects other examples in the literature. For example, there is a similar number in other poverty mapping applications, such as Spain (Molina and Rao, 2010; Molina et al., 2014), where survey data plus model-based census data both at unit level (but no contextual variables) are available.

Despite there being some discrepancies between the sample based and the population based estimator of variability, the main conclusion is that the relative ranking of the variables remains very similar. Using the Spearman correlation it is possible to compare the ordering of the variables using the census data to that for the naïve estimate (5.11) (which uses the survey data only). The Spearman correlation at the commune level is 0.916 with a 95% confidence interval (CI) of (0.77, 0.976), indicating that the variables remain in much the same order with respect to their variability, despite the naïve sample variance (5.11) being biased. The Elbers et al. (2002) approximation method gives a Spearman rank correlation of 0.9252 with a 95% CI (0.781, 0.979); this indicates that the relative order of the variables remains similar using the approximation method compared to the population data. For this data set then, the methods applied to unit record survey data alone (as would be necessary if there was no unit record census data) yield very similar modelling choices to the 'gold standard' when both survey and census unit record data is available.

Aggregation of small areas up to district and province level increases the precision because the sample size is larger, leading to the ordering of the variables in terms of their importance remaining relatively constant. This is indicated by the Spearman correlation of the order of the variables in terms of their importance being 0.897 with a 95% confidence interval of (0.771, 0.954) and 0.922 with a 95% CI of (0.805, 0.975) respectively. The 95% confidence intervals overlap substantially at each level of amalgamation, meaning that (despite the lim-

ited sample size in each of the sampled small areas) the variables retain their relative order for $\{V(\hat{\tau}_{ip}|\hat{\beta}_p)\}^{1/2}$. At each level of aggregation the Spearman correlation for survey data alone is still measuring the relative ordering for the 35 variables included in the model, since aggregation to small area level does not change the number of variables.

Aggregating the sample data to give auxiliary variable mean estimates for fewer small areas does not necessarily remove the overestimation of variance when using sample rather than census data. The sample was not designed with the variables used in the regression model in mind, hence no matter how much aggregation happens the overestimation may never be completely removed. This suggests a Spearman correlation of less than one is a consequence of the sampling scheme. The question of optimal survey design for SAE is discussed for example in Haslett (2012).

## 5.4 Conclusion

The main focus for SAE is to achieve reliable small area predictions, which leads to considerable attention being given to estimating the MSE. However, considerably less attention has been given to determining the auxiliary variables that are to be included in the $\mathbf{X}$ matrix in the linear or generalized linear model which underlies the SAEs. In order to maximize the explained variation at the unit level, models tend to be complicated. The importance of a variable at unit level does not necessarily correspond to the variable being important at the small area level. Standard variable selection techniques such as F- (or t-) tests alone provide limited guidance and a method that also considers variation in the auxiliary variables across small areas, as has been developed in this chapter, is also required. The variable importance measure presented allows one to assess each variable's importance and decide whether it is significantly changing the final small area estimate. Although the diagnostic was fitted to the linear regression model,

it was still shown to be applicable for the non-linear transformation of the data.

For the Cambodian data, I have demonstrated that not all variables that have predictive power in the regression model at the unit level are required for prediction at the small area level. Additionally because the interest is not regression model parameter estimation, SAEs can be relatively robust to reduction in model complexity, as the model may be reducible in terms of the number of parameters without significantly affecting the SAE point estimates. To a lesser extent, the conclusion is the same if the standard errors and the spatial distribution or ranking of the small area estimates are also considered. Furthermore, the exercise of variable ranking and removal can still be achieved when only survey data is available and/or when the model is fitted at area level; the empirical evidence suggests that the relative importance of the variables remains largely unaltered. The more general conclusion based on the empirical evidence is that the diagnostic developed in this paper can provide a useful aid in development of suitable SAE models from unit record survey data, whether or not the census unit record data is available. As well as being applicable to unit level models, the model ranking technique is also relevant for small area estimation methods based on area level models.

## 5.5 Appendix: Derivation of Unbiased Variance Estimator for SAE for a Survey without Census Information

For a given variable $p$, the focus is to take the variability of $\hat{\tau}_{ip}$ across the small areas. For simplicity of exposition an equi-probability selection methods (EPSEM) or self weighting sample designs is considered, but the issues mentioned are more general and can be extended to other designs.

An estimate of $\hat{\tau}_{ip}$ using only survey data can be defined as $\hat{\tau}_{ip} = \hat{\beta}_p \hat{\bar{x}}_{ip}$, however the

variability of $\hat{\tau}_{ip}$ taken across the sampled small areas is a biased estimator of the population variability, as shown

$$V(\hat{\bar{x}}_{ip}|\hat{\beta}_p) \neq V(\bar{x}_{ip}|\hat{\beta}_p) \tag{5.13}$$

where

$$V(\hat{\bar{x}}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}\sum_{i\in s}^{I_s}(\hat{\bar{x}}_{ip} - \hat{\bar{x}}_{.p})^2 \tag{5.14}$$

and $\hat{\bar{x}}_{ip}$ is the estimated mean for each of the $p$ variables in each of the $I_s$ sampled small areas. The notation '.' indicates the mean of the variable, for example $\hat{\bar{x}}_{.p}$ is the population mean for variable $i$ using only the sample data; the sum is taken over the $I_s$ small areas included in the sample. These sampled small area level values are estimable from the survey. However (5.14) is a biased estimate of the corresponding population variance as its expectation is not the population variability which is:

$$V(\bar{x}_{ip}|\hat{\beta}_p) = \frac{1}{I - 1}\sum_{i=1}^{I}(\bar{x}_{ip} - \bar{x}_{.p})^2 \tag{5.15}$$

This includes all $I$ small areas from the population. $\bar{x}_{ip}$ is the small area population mean for each variable within each small area, obtained from all unit level census records. The census level mean for each variable $p$ is $\bar{x}_{.p}$, although these are only available if there is unit level census data.

In order to determine the bias of the variance in (5.14), we expand to include the terms in (5.15) and determine the additional non-zero terms.

$$V(\hat{\bar{x}}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}\sum_{i\in s}^{I_s}[(\bar{x}_{ip} - \tilde{\bar{x}}_{.p}) + (\tilde{\bar{x}}_{.p} - \hat{\bar{x}}_{.p}) + (\hat{\bar{x}}_{ip} - \bar{x}_{ip})]^2 \tag{5.16}$$

where $\tilde{\bar{x}}_{.p}$ is defined as the true mean taken over $I_s$ sampled areas.

This is equivalent to (5.14). Now an unbiased estimator of the variance is

$$V(\bar{x}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}\sum_{i\in s}^{I}(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})^2 \tag{5.17}$$

this corresponds to the square of the first term in (5.16). Further expanding the (5.16) gives

$$\begin{aligned} V(\hat{\bar{x}}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}[&\sum_{i\in s}^{I_s}(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})^2 + I_s(\bar{\tilde{x}}_{.p} - \hat{\bar{x}}_{.p})^2 + \sum_{i\in s}^{I_s}(\hat{\bar{x}}_{ip} - \bar{x}_{ip})^2 \\ &+2(\bar{\tilde{x}}_{.p} - \hat{\bar{x}}_{.p})\sum_{i\in s}^{I_s}(\bar{x}_{ip} - \bar{\tilde{x}}_{.p}) + 2\sum_{i\in s}^{I_s}(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})(\hat{\bar{x}}_{ip} - \bar{x}_{ip}) + 2(\bar{\tilde{x}}_{.p} - \hat{\bar{x}}_{.p})\sum_{i\in s}^{I_s}(\hat{\bar{x}}_{ip} - \bar{x}_{ip})] \end{aligned} \tag{5.18}$$

Note that $\sum_{i\in s}^{I_s}(\bar{x}_{ip} - \bar{\tilde{x}}_{.p}) = 0$ and $\sum_{i\in s}^{I_s}(\hat{\bar{x}}_{ip} - \bar{x}_{ip}) = -I_s(\bar{\tilde{x}}_{.p} - \hat{\bar{x}}_{.p})$ so that

$$V(\hat{\bar{x}}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}\sum_{i\in s}^{I_s}[(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})^2 + (\hat{\bar{x}}_{ip} - \bar{x}_{ip})^2 - (\bar{\tilde{x}}_{.p} - \hat{\bar{x}}_{.p})^2 + 2(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})(\hat{\bar{x}}_{ip} - \bar{x}_{ip})] \tag{5.19}$$

An unbiased estimator of the variance would be

$$V(\hat{\bar{x}}_{ip}|\hat{\beta}_p) = \frac{1}{I_s - 1}\sum_{i\in s}^{I_s}(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})^2. \tag{5.20}$$

However there are three terms in (5.19) that are inflating the variance, these being:

i) $\frac{I_s}{I_s - 1}(\bar{\tilde{x}}_{.p} - \hat{\bar{x}}_{.p})^2$

ii) $\frac{1}{I_s - 1}\sum_{i\in s}^{I_s}(\hat{\bar{x}}_{ip} - \bar{x}_{ip})^2$

iii)$\frac{1}{I_s - 1}\sum_{i\in s}^{I_s}2(\bar{x}_{ip} - \bar{\tilde{x}}_{.p})(\hat{\bar{x}}_{ip} - \bar{x}_{ip})$

where

i) For a particular variable $p$, this component is relatively small, as it is the squared difference in the overall mean of the variable from the sample and the census data.

ii) Is the difference between the small area mean for a variable measured in the survey and the

census. If the sample in each small area is relatively small, this component can be quite large. Amalgamating the small areas in order to create larger samples in each small area can decrease this.

iii) This component can either be positive or negative.

In order to gain an unbiased estimate of the variance, the expected values of the terms that inflate the variance estimator would need to be estimated unbiasedly and subtracted from the sample variance. However an unbiased estimator for the variance $\bar{x}_{ip}$ across the small areas using only the survey data is not generally possible. Although as long as the additional terms are small or relatively constant it should not be important in the exercise of ranking the variables by relative importance for small area estimation. The examination of the Cambodia data indicates this is a plausible situation.

## 5.6 Appendix 2

[1]In Table 5.5: hhsXS3 is hhsizeXS3, rXS3* is roof_cXS3, nXS3* is numroomXS3 and mXS3* is motorbikeXS3

Table 5.1: VIM of the variables in the Cambodian poverty rate model.

| var removed | $\hat{\beta}_p$ | t-stat | sd($\bar{x}_{ip}$) | $sd(\hat{\tau}_{ip}\|\hat{\beta}_p)$ |
|---|---|---|---|---|
| radio | 0.019 | 2.19 | 0.1121 | 0.0021 |
| phone | 0.125 | 2.31 | 0.0176 | 0.0022 |
| roof_cXS3 | 0.159 | 3.17 | 0.0162 | 0.0026 |
| pkids06 | -0.109 | -4.31 | 0.0289 | 0.0032 |
| rfree | -0.120 | -4.97 | 0.0312 | 0.0038 |
| computer | 0.101 | 3.95 | 0.0647 | 0.0065 |
| motorbikeXS3 | 0.036 | 2.30 | 0.2135 | 0.0077 |
| floor_c | 0.024 | 3.76 | 0.2957 | 0.0070 |
| floor_t | 0.103 | 4.55 | 0.1191 | 0.0123 |
| resplus_e | 0.233 | 2.96 | 0.0454 | 0.0106 |
| numroomXS3 | -0.029 | -1.84 | 0.4426 | 0.0128 |
| roof_m | 0.056 | 3.90 | 0.2020 | 0.0113 |
| roof_c | 0.060 | 1.87 | 0.1057 | 0.0064 |
| roof_t | 0.112 | 6.70 | 0.2157 | 0.0242 |
| floor_s | 0.418 | 2.52 | 0.0326 | 0.0136 |
| electric | 0.051 | 2.37 | 0.2731 | 0.0140 |
| pseced | 0.084 | 4.30 | 0.1270 | 0.0107 |
| notoilet | -0.050 | -3.46 | 0.2440 | 0.0122 |
| wall_b | -0.060 | -4.86 | 0.2053 | 0.0123 |
| car | 0.265 | 10.54 | 0.0745 | 0.0198 |
| boat_e | 0.152 | 3.74 | 0.2232 | 0.0340 |
| numroom | 0.096 | 7.81 | 0.2260 | 0.0217 |
| plnmount | -0.067 | -2.69 | 0.3972 | 0.0267 |
| tonlesap | -0.060 | -3.53 | 0.4507 | 0.0269 |
| plit | 0.114 | 6.05 | 0.1389 | 0.0158 |
| plit_e | -0.453 | -3.32 | 0.1384 | 0.0627 |
| h_lit_e | 0.322 | 2.78 | 0.1400 | 0.0451 |
| motorbike | 0.085 | 6.27 | 0.2484 | 0.0211 |
| hhsizeXS3 | 0.032 | 5.56 | 1.5104 | 0.0479 |
| reg3 | -0.153 | -3.52 | 0.3104 | 0.0475 |
| hhsize | -0.034 | -4.01 | 0.4326 | 0.0149 |
| tv | 0.075 | 7.87 | 0.2517 | 0.0188 |
| lnhhsize | -0.547 | -16.11 | 0.0934 | 0.0511 |
| cellphone | 0.123 | 16.96 | 0.3942 | 0.0485 |
| cellphone_e | 0.165 | 6.77 | 0.4649 | 0.0765 |
| constant | 9.310 | 141.10 | | |

Table 5.2: Correlation matrix of ln_exp and the regression covariates part 1.

| | ln_exp | hhsize | lnhhsz | pkids06 | plit | pseced | notoilet | numroom | rfree |
|---|---|---|---|---|---|---|---|---|---|
| ln_exp | 1.000 | | | | | | | | |
| hhsize | -0.282 | 1.000 | | | | | | | |
| lnhhsz | -0.303 | 0.948 | 1.000 | | | | | | |
| pkids06 | -0.214 | 0.071 | 0.129 | 1.000 | | | | | |
| plit | 0.343 | 0.073 | 0.089 | -0.292 | 1.000 | | | | |
| pseced | 0.443 | 0.035 | 0.045 | -0.190 | 0.561 | 1.000 | | | |
| notoilet | -0.451 | -0.048 | -0.042 | 0.128 | -0.327 | -0.414 | 1.000 | | |
| numroom | 0.382 | 0.182 | 0.163 | -0.135 | 0.236 | 0.309 | -0.339 | 1.000 | |
| rfree | -0.004 | -0.074 | -0.077 | 0.078 | -0.020 | -0.004 | -0.027 | 0.016 | 1.000 |
| car | 0.352 | 0.081 | 0.075 | -0.037 | 0.143 | 0.218 | -0.211 | 0.313 | -0.021 |
| cellphone | 0.548 | 0.194 | 0.190 | -0.119 | 0.362 | 0.481 | -0.485 | 0.427 | -0.020 |
| computer | 0.333 | 0.074 | 0.066 | -0.069 | 0.159 | 0.274 | -0.228 | 0.279 | 0.008 |
| electric | 0.502 | 0.031 | 0.021 | -0.097 | 0.295 | 0.406 | -0.548 | 0.326 | 0.067 |
| motorbike | 0.427 | 0.242 | 0.248 | -0.100 | 0.330 | 0.417 | -0.384 | 0.376 | -0.022 |
| phone | 0.112 | 0.019 | 0.018 | -0.023 | 0.057 | 0.091 | -0.076 | 0.079 | 0.017 |
| radio | 0.098 | 0.036 | 0.025 | -0.139 | 0.100 | 0.105 | -0.092 | 0.099 | -0.048 |
| tv | 0.402 | 0.196 | 0.199 | -0.126 | 0.323 | 0.346 | -0.367 | 0.358 | -0.064 |
| floor_t | 0.438 | 0.028 | 0.021 | -0.055 | 0.206 | 0.321 | -0.360 | 0.341 | 0.022 |
| floor_c | 0.171 | -0.016 | -0.018 | -0.004 | 0.111 | 0.123 | -0.199 | 0.057 | 0.059 |
| floor_s | 0.033 | 0.010 | 0.009 | -0.011 | 0.010 | 0.019 | -0.032 | 0.026 | -0.005 |
| roof_t | 0.080 | 0.075 | 0.080 | -0.097 | 0.133 | 0.115 | -0.065 | 0.178 | -0.006 |
| roof_c | 0.310 | 0.041 | 0.030 | -0.053 | 0.136 | 0.203 | -0.221 | 0.236 | 0.010 |
| roof_m | -0.013 | -0.055 | -0.051 | 0.042 | 0.002 | -0.020 | -0.031 | -0.121 | 0.015 |
| wall_b | -0.335 | -0.081 | -0.081 | 0.098 | -0.255 | -0.290 | 0.351 | -0.290 | -0.040 |
| boat_e | -0.037 | 0.023 | 0.021 | -0.007 | -0.036 | -0.062 | 0.065 | -0.030 | -0.004 |
| cellphone_e | 0.581 | 0.042 | 0.028 | -0.100 | 0.317 | 0.433 | -0.538 | 0.368 | 0.034 |
| h_lit_e | 0.330 | -0.019 | -0.022 | -0.084 | 0.346 | 0.349 | -0.348 | 0.165 | 0.017 |
| plit_e | 0.357 | -0.019 | -0.023 | -0.112 | 0.383 | 0.388 | -0.397 | 0.192 | 0.003 |
| resplus_e | 0.291 | -0.001 | 0.000 | -0.035 | 0.142 | 0.184 | -0.252 | 0.136 | 0.006 |
| reg3 | -0.461 | -0.028 | -0.021 | 0.074 | -0.268 | -0.361 | 0.477 | -0.303 | -0.037 |
| tonlesap | -0.117 | 0.000 | 0.002 | 0.001 | -0.034 | -0.067 | -0.008 | -0.033 | -0.005 |
| plnmount | -0.098 | 0.015 | 0.015 | 0.046 | -0.107 | -0.093 | 0.114 | -0.009 | 0.019 |
| hhsizeXS3 | -0.525 | 0.563 | 0.545 | 0.097 | -0.154 | -0.256 | 0.337 | -0.158 | -0.065 |
| roof_cXS3 | 0.025 | 0.024 | 0.022 | 0.015 | -0.003 | -0.006 | -0.002 | 0.027 | 0.041 |
| numroomXS3 | -0.208 | 0.067 | 0.067 | -0.021 | -0.086 | -0.149 | 0.200 | 0.298 | -0.010 |
| motorbikeXS3 | 0.068 | 0.158 | 0.174 | -0.034 | 0.139 | 0.108 | -0.051 | 0.068 | -0.036 |

Table 5.3: Correlation matrix of ln_exp and the regression covariates part 2.

| | car | cellphone | computer | electric | motorbike | phone | radio | tv | floor_t |
|---|---|---|---|---|---|---|---|---|---|
| car | 1.000 | | | | | | | | |
| cellphone | 0.356 | 1.000 | | | | | | | |
| computer | 0.351 | 0.395 | 1.000 | | | | | | |
| electric | 0.243 | 0.534 | 0.292 | 1.000 | | | | | |
| motorbike | 0.223 | 0.592 | 0.307 | 0.387 | 1.000 | | | | |
| phone | 0.146 | 0.098 | 0.098 | 0.092 | 0.091 | 1.000 | | | |
| radio | 0.063 | 0.088 | 0.041 | 0.025 | 0.101 | 0.033 | 1.000 | | |
| tv | 0.282 | 0.489 | 0.253 | 0.377 | 0.463 | 0.096 | 0.102 | 1.000 | |
| floor_t | 0.327 | 0.453 | 0.380 | 0.451 | 0.310 | 0.075 | 0.007 | 0.293 | 1.000 |
| floor_c | 0.039 | 0.155 | 0.028 | 0.236 | 0.124 | 0.025 | 0.016 | 0.120 | -0.094 |
| floor_s | 0.024 | 0.025 | 0.013 | 0.028 | 0.015 | -0.003 | -0.003 | 0.010 | -0.008 |
| roof_t | 0.018 | 0.048 | -0.018 | -0.054 | 0.123 | 0.009 | 0.082 | 0.164 | -0.076 |
| roof_c | 0.224 | 0.328 | 0.299 | 0.306 | 0.213 | 0.061 | 0.004 | 0.194 | 0.498 |
| roof_m | -0.056 | -0.016 | -0.054 | 0.087 | -0.026 | -0.007 | -0.026 | -0.022 | -0.046 |
| wall_b | -0.129 | -0.329 | -0.130 | -0.320 | -0.330 | -0.055 | -0.091 | -0.303 | -0.211 |
| boat_e | -0.041 | -0.050 | -0.051 | -0.102 | -0.084 | -0.018 | 0.015 | -0.023 | -0.078 |
| cellphone_e | 0.328 | 0.602 | 0.410 | 0.736 | 0.422 | 0.087 | 0.039 | 0.396 | 0.596 |
| h_lit_e | 0.147 | 0.318 | 0.168 | 0.375 | 0.227 | 0.036 | 0.059 | 0.287 | 0.252 |
| plit_e | 0.161 | 0.356 | 0.189 | 0.422 | 0.255 | 0.054 | 0.066 | 0.318 | 0.283 |
| resplus_e | 0.143 | 0.262 | 0.158 | 0.322 | 0.161 | 0.084 | 0.002 | 0.163 | 0.255 |
| reg3 | -0.242 | -0.491 | -0.319 | -0.681 | -0.333 | -0.086 | -0.018 | -0.319 | -0.484 |
| tonlesap | -0.037 | -0.080 | -0.065 | -0.045 | -0.088 | -0.019 | 0.011 | -0.096 | -0.094 |
| plnmount | -0.026 | -0.085 | -0.052 | -0.099 | -0.044 | -0.014 | -0.019 | -0.121 | -0.093 |
| hhsizeXS3 | -0.167 | -0.292 | -0.234 | -0.503 | -0.144 | -0.064 | -0.007 | -0.144 | -0.358 |
| roof_cXS3 | 0.014 | 0.017 | 0.001 | 0.021 | 0.018 | 0.000 | 0.000 | 0.002 | 0.019 |
| numroomXS3 | -0.106 | -0.223 | -0.210 | -0.413 | -0.101 | -0.051 | 0.045 | -0.092 | -0.299 |
| motorbikeXS3 | -0.015 | 0.148 | -0.077 | -0.096 | 0.623 | 0.015 | 0.081 | 0.190 | -0.109 |

Table 5.4: Correlation matrix of ln_exp and the regression covariates part 3.

| | floor_c | floor_s | roof_t | roof_c | roof_m | wall_b | boat_e | cellphone_e | h_lit_e |
|---|---|---|---|---|---|---|---|---|---|
| floor_c | 1.000 | | | | | | | | |
| floor_s | -0.008 | 1.000 | | | | | | | |
| roof_t | -0.083 | 0.002 | 1.000 | | | | | | |
| roof_c | 0.042 | 0.009 | -0.163 | 1.000 | | | | | |
| roof_m | 0.147 | 0.003 | -0.658 | -0.203 | 1.000 | | | | |
| wall_b | -0.172 | -0.012 | -0.284 | -0.140 | 0.032 | 1.000 | | | |
| boat_e | -0.063 | -0.004 | 0.001 | -0.058 | 0.006 | 0.083 | 1.000 | | |
| cellphone_e | 0.206 | 0.025 | -0.099 | 0.480 | 0.049 | -0.317 | -0.100 | 1.000 | |
| h_lit_e | 0.117 | 0.022 | 0.065 | 0.165 | 0.036 | -0.164 | -0.050 | 0.475 | 1.000 |
| plit_e | 0.124 | 0.021 | 0.086 | 0.188 | 0.033 | -0.190 | -0.080 | 0.524 | 0.922 |
| resplus_e | 0.113 | -0.010 | -0.080 | 0.218 | 0.042 | -0.140 | -0.050 | 0.445 | 0.196 |
| reg3 | -0.197 | -0.035 | 0.120 | -0.339 | -0.100 | 0.265 | 0.102 | -0.771 | -0.374 |
| tonlesap | -0.034 | 0.004 | -0.117 | -0.109 | 0.107 | 0.049 | 0.013 | -0.105 | -0.146 |
| plnmount | -0.018 | -0.010 | -0.033 | -0.029 | -0.031 | -0.037 | -0.026 | -0.112 | -0.167 |
| hhsizeXS3 | -0.156 | -0.028 | 0.139 | -0.245 | -0.103 | 0.147 | 0.091 | -0.569 | -0.301 |
| roof_cXS3 | 0.057 | -0.003 | -0.069 | 0.424 | -0.086 | -0.032 | -0.017 | 0.003 | -0.017 |
| numroomXS3 | -0.100 | -0.023 | 0.239 | -0.223 | -0.130 | 0.020 | 0.084 | -0.502 | -0.238 |
| motorbikeXS3 | 0.022 | -0.012 | 0.205 | -0.095 | -0.044 | -0.163 | -0.014 | -0.170 | -0.035 |

Table 5.5: Correlation matrix of ln_exp and the regression covariates part 4.

| | plit_e | resplus_e | reg3 | tonlesap | plnmount | hhs$\tilde{X}$S3 [1] | r$\tilde{X}$S3* | nXS3* | mXS3* |
|---|---|---|---|---|---|---|---|---|---|
| plit_e | 1.000 | | | | | | | | |
| resplus_e | 0.208 | 1.000 | | | | | | | |
| reg3 | -0.412 | -0.319 | 1.000 | | | | | | |
| tonlesap | -0.110 | -0.018 | 0.059 | 1.000 | | | | | |
| plnmount | -0.240 | -0.046 | 0.064 | -0.244 | 1.000 | | | | |
| hhsizeXS3 | -0.332 | -0.241 | 0.747 | 0.054 | 0.066 | 1.000 | | | |
| roof_cXS3 | -0.033 | 0.005 | 0.048 | -0.007 | 0.059 | 0.055 | 1.000 | | |
| numroomXS3 | -0.257 | -0.215 | 0.711 | 0.042 | 0.068 | 0.597 | 0.074 | 1.000 | |
| motorbikeXS3 | -0.034 | -0.067 | 0.335 | -0.024 | 0.013 | 0.377 | 0.057 | 0.402 | 1.000 |

# Chapter 6

# An Influence Diagnostic for SAE

Small area estimation uses statistical models to improve the precision of survey-based estimates for small subdomains of the target population. Considerable attention has been given to estimating the mean square error of small area estimates, but much less attention has been given to identifying any unusual small area estimates and checking that the model and data sources are correct, at least in the context of unit record data with small area estimates involving aggregation in terms of thousands or tens of thousands of observations. The methodology developed in this chapter was motivated by the need to investigate and remedy an anomalous small area estimate of wasting during an undernutrition mapping exercise in Nepal. I propose an influence diagnostic for small area estimation that focuses on the combinations of regression parameters and auxiliary data that are most important for a particular small area estimate. The theoretical justification for the proposed diagnostic here is primarily for linear models but may be shown to be extendible to quantify the relative effect of each auxiliary variable in determining the predictions from non-linear outcomes such as poverty and undernutrition rates, based on linear models. This allows the identification of any variables within an area that have a larger than expected influence on the small area estimate for that area, and this highlights possible errors,

which need to be checked and if necessary corrected.

## 6.1 Introduction

In general, regression models are often fitted in order to make predictions for future unspecified data. However in certain poverty mapping applications of SAE, where there is unit record survey and census data, the purpose of the survey model is very specific: the model is applied to another set of the same known covariates from a census and then later unit level predictions are aggregated to small area level estimates. Regression diagnostics are used to check whether the model assumptions hold, as well as identifying the influence that an individual or subset of data has on the outcome of interest. Diagnostics in general tend to focus on the 'training data' to which the model is fitted, as outlined in Section 2.5. Some examples on this are in SAE where Pfeffermann (2013) outlines recent developments in accounting for measurement errors in covariates, as well as treatments for outliers. Chambers et al. (2014) focuses on making SAE robust to model outliers which are present in the training data and Baldermann et al. (2018) makes robust small area estimation in the presence of spatial non-stationarity. However, current methods tend to apply only to the effect the data has on the fitted model, rather than the final small area predictions. Despite rapid advancements in small area model fitting techniques, diagnostics checking the validity of these models and identifying outlying or unusual small area estimates (rather than model outliers) have largely been neglected. When the estimate for a particular small area is felt to be unusual, for example based on expert opinion, it would be useful to explore which variables and observations appear to be the causes, so that possible remedies can be sought. Experience of this situation in the small area estimation of undernutrition indicators in Nepal (Haslett et al., 2014a) provided the motivation to investigate diagnostics to identify influential or unusual auxiliary observations affecting the small area estimates. An auxiliary

variable is a variable that is known for every unit of the population but is not the variable of interest, but is rather used to improve the sampling plan or enhance estimation of the variable of interest (Lavrakas, 2008).

This chapter is organised as follows: Section 6.2 outlines the proposed methodology for identifying influential observations or small areas by considering both the regression parameters as well as the auxiliary information. Section 6.3 gives details of the application to undernutrition mapping in Nepal from which the example arose. Section 6.3.1 extends the diagnostic to show it can be used as a general tool to check for any other potential errors in mapping exercises. Finally conclusions are drawn in Section 6.4.

## 6.2 Methodology

The proposed deletion diagnostic for the small area estimates incorporates the multiple data sources used in SAE, and focuses on how the predicted area level response is affected by perturbations in the data. More specifically the focus is the difference between the small area level mean and the population level mean (or a localised mean at some higher level), for each auxiliary variable in the model and the influence each of these differences has on the SAE estimate for that area.

The diagnostic was originally developed for situations where the model is fitted using survey data and predictions are made using census data. It is however still useful when only survey data is available.

The situation when census and survey data are both available at unit record level is discussed first. In this case, the survey data is used to estimate the regression parameters. The fitting methodology should incorporate the survey design; see for example Section 3.3.1 for

more details. The model is then used to predict each unit record in the census data, and then these unit-level predictions are aggregated up to small area level. The underlying survey based model is

$$Y = \mathbf{X}\beta + \varepsilon \tag{6.1}$$

where $\mathbf{X}$, the auxiliary variables, include both unit level data common to both the survey and census, as well as contextual variables at psu or finer level. The parameter vector $\beta$ is of length $p$, where this is the number of regression parameters in the model and $\varepsilon$ is the structured error term as in the nested regression model from Section 2.2.1. As in the previous chapter an individual unit $k$ within a small area $i$ is given by

$$y_{ik} = \mathbf{x}'_{ik}\beta + u_i + e_{ik}; \qquad i = 1,...,I; \quad k = 1,...,n_i \tag{6.2}$$

where $\mathbf{x}_{ik} = (x_{ik1},...,x_{ikp},...,x_{ikP})'$ is a vector of auxiliary information at unit level, $n_i$ is the sample size in the $i^{th}$ small area and the unit level error is $e_{ik}$. It is usual but not necessary to assume that the survey errors $\{\tilde{e}_{ik}\}$ and the random effects $\{u_i\}$ are normally distributed and independent error terms with an expected mean of zero. When the primary sample units (psu) are also the defined 'small areas', then $u_i$ should capture this unexplained variation at this level, however an additional psu level error term may be required if the small areas contain several psus. Conditional on $i$, $\tilde{e}_{ik}$ and $u_i$ may or may not have zero mean; ensuring approximately zero mean is an important role for the contextual variables, see for example Haslett et al. (2014b); Molina and Rao (2010) considers the case of non-zero mean $v_i$ given $i$.

The fitted model is applied to the census in order to produce 'synthetic' small area estimates, this is shown as:

$$\hat{\bar{y}}_i = \bar{x}'_{ip}\hat{\beta}_p \quad i = 1,...,I \quad p = 1,...,P. \tag{6.3}$$

which is supplementary to the direct estimate $\tilde{y}_i$ from the survey data for that area. Here $\bar{x}_{ip}$ is the mean for small area $i$ for variable $p$. If estimated standard errors are required then (6.3) can, for example, be supplemented by bootstrap residuals $B$ times, for some $B$. For more detail see Section 3.3.1.

When composite estimation is possible, that is when there is a sample in each small area, the SAE is a weighted mean of the model prediction and the direct estimate. The weighting depends on the relative size of the variance of the direct estimator and the error variance of the model prediction at area level. In this specific application, for each small area there are contextual variables (at cluster or finer level) but sometimes little (or no) sample unit record data. Moreover the inclusion of psu (cluster) or finer level covariates as contextual variables can greatly reduces the area level error variance from the model-based prediction so that the contribution of the direct estimate can become negligible, see for example (2.5). We therefore focus on the model-based part.

As with the more usual regression influence diagnostics, see for example Cook and Weisberg (1982), a large influence statistic indicates a variable $p$ that is having a large effect on the estimate in small area $i$. For a given small area $i$, the contribution of variable $p$ to the small area estimate will be treated as unusual if the area level mean for that variable differs markedly from the population mean, and has a relatively large and statistically significant regression parameter. It is the product of these two terms that is important, not the regression coefficient alone. Such variables can be further investigated to determine if the final small area estimate is a true reflection of the variable of interest or if there is an error in the data or any other anomaly in variable $p$, in which case an appropriate remedy needs to be considered and implemented.

From (6.3) an influence statistic can be generated for each variable $p$ in each small area $i$ by combining the regression coefficient with the difference between the small area and population mean, this is calculated as:

$$\phi_{ip} = (\bar{x}_{ip} - \bar{\bar{x}}_{(i)p})\hat{\beta}_p \tag{6.4}$$

where $\bar{\bar{x}}_{(i)p}$ is the weighted population mean for variable $p$, excluding small area $i$ and $\bar{x}_{ip}$ is the mean for the $i^{th}$ small area. This results in an $I \times P$ matrix of influence diagnostics. It may be thought that an influence diagnostic can be formed by just taking the difference between the small area mean and the weighted population mean $(\bar{x}_{ip} - \bar{\bar{x}}_{(i)p})$. However the regression coefficient $\hat{\beta}_p$ is needed as it is contributing to the overall effect of the response. Note that $\phi_{ip}$ is scale free: if $x_p$ is rescaled then $\beta_p$ adjusts so that $\phi_{ip}$ remains the same. Note also that $\phi_{ip} = \hat{\tau}_{ip} - \hat{\tau}_{(i)p}$ from (5.8).

In situations where there are a large number of variables and small areas this can result in a large number of diagnostics to analyse, for example in the case of measuring the poverty rate in Cambodia there were 1621 small areas and 35 variables in the model, resulting in 56,735 influence diagnostics. However these influence diagnostics can be examined graphically to see if any are unusually large. It might be thought useful to conduct numerical analysis by having a threshold to determine if a particular value is influential. In this situation the influence diagnostic could be standardized based on $i$ or $p$ by taking the inverse of the square root of $var[(\bar{x}_{ip} - \bar{\bar{x}}_{(i)p})\hat{\beta}_p]$, this can be done using the sample variance. However Fox (1991) argues that numerical thresholds should be used with caution, and that graphical displays should instead be used to assess which observations need further examination. Alternatively, if some of the small area estimates are judged to be unusual (perhaps after a validation study), the diagnostics can help to identify which variables are driving the abnormalities.

Further, even if there is a non-linear transformation of model predictions, for example from the standardized weight for height (WHZ) to the wasting rate (which is an indicator variable that has been aggregated), the diagnostic (6.4) still provides a useful indicator. If the linear

predictor changes by a large amount, then it is likely that the aggregated non-linear transforms will change appreciably too. The evidence from the case study in Section 6.3 supports this.

If the auxiliary data are only available from the survey, (6.4) can be estimated using survey data. The small area level means and the population means for each variable are then replaced by their survey weighted means $\hat{\bar{x}}_{ip}$ and $\hat{\bar{\bar{x}}}_{(i)p}$ respectively. It is useful in this situation if all small areas are sampled in the survey, because otherwise these estimates would be unobtainable for small areas. Moreover $\hat{\bar{x}}_{ip}$, will then be subject to sampling error, making the diagnostic less useful unless the sample sizes in small areas are reasonably large.

For the area level model, data are already aggregated to their area level means. In the situation when there are area level census or administrative variables available, (6.4) can still be applied. This is similar to the situation when just unit record survey data is available, the principal difference being that the area level data is already in aggregated form. However area level models commonly use covariates obtained from census and administrative data, in which case sampling uncertainty in the area-level means is not an issue.

Small areas in close geographical proximity tend to have similar characteristics. It is possible that a small area mean for a particular area is not unusual when compared to the population as a whole, but is unusual compared to the small areas within its region (which is the reason diagnostics are better fitted and assessed before applying any spatial smoothing). It may be useful therefore to adapt the influence diagnostic (6.4) so that, instead of using the population mean ($\bar{\bar{x}}_{(i)p}$), a localised mean such as at regional or district level is used in its place. The resulting influence diagnostic for variable $p$ and small area $i$ becomes

$$\phi_{ap} = (\bar{x}_{ip} - \bar{\bar{x}}_{a(i)p})\hat{\beta}_p \tag{6.5}$$

where $a$ is the chosen geographical level in which the local small areas are being compared

and $\bar{\bar{x}}_{a(i)p}$ is the mean for that geographical level excluding small area $i$. Defining the small area influence statistic at a more localised level of aggregation identifies small areas within the defined area that are behaving differently from other areas in the region. Care must be taken when analysing at a localised level, for example, the anomalous area in a region could be the only urban part of an otherwise rural region so it may appear unusual. So it is important to check anomalous results carefully, using local knowledge if available (see Section 6.3.1 for examples). Equation (6.5) can be applied to unit or area level data, and can be adapted for survey data by using the weighted survey means as estimates.

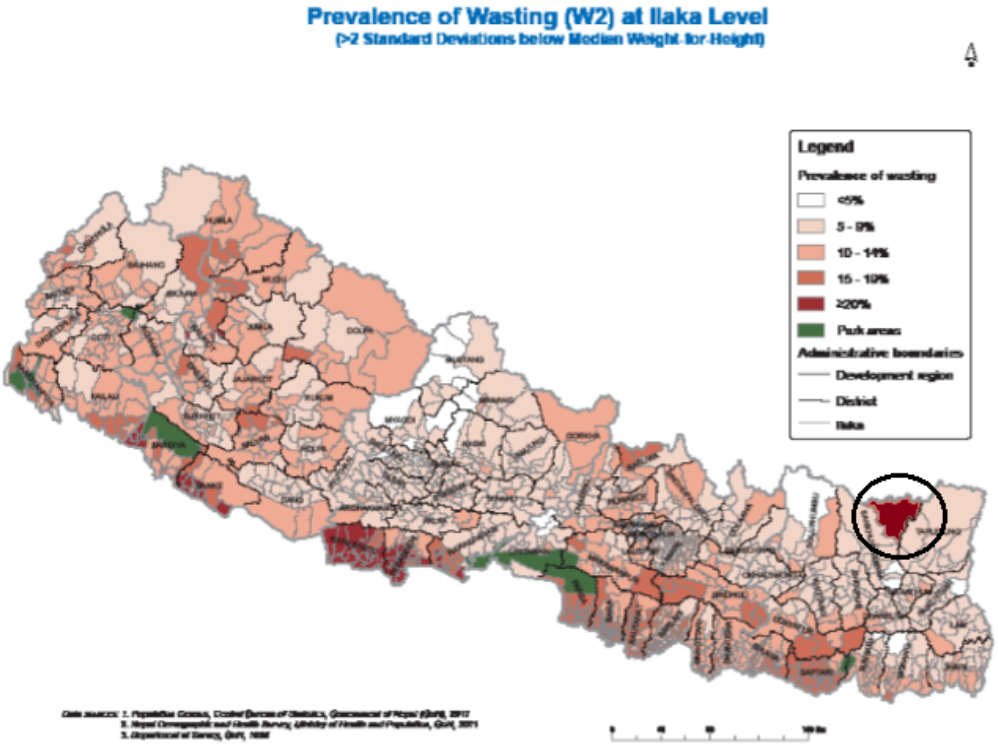
## 6.3 Application to the Wasting Rate in Nepal

The inspiration for this diagnostic technique came during a validation exercise using local expert knowledge for a particularly unusual wasting rate in Nepal. Wasting is a measure of acute malnutrition of a child/children, often linked to diarrhoea-causing diseases, and is based on the standardized weight for height of a child (WHZ). This in turn defines the wasting rate, which is the proportion of the population of children under 5 who have a standardised weight for height of less than -2 compared to the reference population, (as outlined in Section 3.1.2). The Nepalese data source used in this application is described in Chapter 4.

Figure 6.1 displays a map of the small area estimates of wasting rates in Nepal at ilaka level; see Section 4.2.3, for more details. The grey lines divide the small areas (ilakas) and the black lines divide the districts. In this particular map, the darker the shade of pink, the higher the wasting rate in the small area.

During a validation exercise using local experts, ilaka 901 in the district of Sankhuwasabha was flagged as suspicious due to the very high estimated wasting rate of 44%. This small area is located in the north east side of Nepal and is circled in Figure 6.1. The high wasting rate

Figure 6.1: Small area estimates of the prevalence of Wasting in Nepal.




would be regarded as a possible humanitarian crisis if it were true that 44% of children who were under five in the area had a low weight for height. Initially it was recommended that immediate aid would need to be distributed to remedy the problem. But before funding and feeding programmes were set up in the small area it was important to check whether the estimated wasting rate was reliable. The small area is logistically very difficult to access, as it takes two days to walk there from the nearest road. This made field verification rather costly and time consuming, so it was desirable to come up with a diagnostic to assess the reliability of the estimate. A further aspect that made the wasting rate in this particular small area seem unusual was that it had a relatively high rate compared to the surrounding small areas in the same district, which all had wasting rates below 14%. Because of the geographical isolation of this ilaka, field verification was difficult. Therefore diagnostic techniques became especially

important in determining which auxiliary variables were causing the small area to have such a high estimated wasting rate.

There was no survey data collected from this small area and the estimated wasting rate was generated by applying the regression model to the census data and making predictions using the Elbers et al. (2003) method; this is described in Section 3.3.1. The unusually high rate could be a true reflection of the nutrition level of the small area or there could be a variable which is erroneously causing the unusually large wasting. Hence the first step is to investigate if there is a particular variable or variables that are driving this deviation.

A high wasting rate implies a low prediction for the WHZ score as can be seen from (3.3). A variable, $p$, causing this would correspond to a negative value of $\phi_{ip}$ in (6.4). An unusual estimate does not necessarily mean there is an error in the data or the model. It could instead be a reflection of the current situation in the small area, perhaps due to localised food shortage or crop failure. The diagnostic (6.4) is used to aid the assessment of this situation.

Table 6.1 displays the estimated regression parameter and associated standard error for each variable included in the initial model, where these were fitted using the survey data. The population mean $(\bar{\bar{x}}_p)$ for each variable is then listed followed by the mean of each of the variables for ilaka 901 $(\bar{x}_{ip})$; here $i = 901$. Column six $(\phi_{901})$ displays the global deletion diagnostics, where the mean of each particular variable in ilaka 901 is compared to overall mean of the variable for the country excluding itself. Column seven $(\phi_{a,901})$ displays the localised deletion diagnostic at the district level, where ilaka 901 is compared to the mean of the remaining ten small areas in the district. These final two columns will help identify the variables that are causing the average child's WHZ in small area 901 to increase or decrease relative to the rest of the country and the rest of the district respectively. This in turn influences the estimated wasting rate of the small area.

Table 6.1: Fitted model and deletion diagnostics for ilaka 901 in Nepal.

| Variable | $\hat{\beta}$ | $se(\hat{\beta})$ | $\bar{\bar{x}}$ | $\bar{x}_{901}$ | $\phi_{901}$ | $\phi_{a,901}$ |
|---|---|---|---|---|---|---|
| ageyr23 | -0.129 | 0.056 | 0.372 | 0.358 | 0.002 | 0.001 |
| girl | 0.108 | 0.047 | 0.490 | 0.474 | -0.002 | -0.002 |
| terai | 0.438 | 0.082 | 0.333 | 0.000 | -0.146 | 0.000 |
| wat_cwell | 0.394 | 0.173 | 0.019 | 0.001 | -0.007 | -0.005 |
| hage2 | -0.153 | 0.058 | 0.324 | 0.432 | -0.017 | -0.012 |
| flr_con | 0.378 | 0.100 | 0.039 | 0.000 | -0.015 | 0.000 |
| wall_wood | 1.341 | 0.340 | 0.043 | 0.040 | -0.004 | 0.015 |
| wall_bambo | 1.216 | 0.316 | 0.168 | 0.200 | 0.039 | -0.071 |
| wall_brk | 1.211 | 0.314 | 0.762 | 0.492 | -0.327 | -0.256 |
| Wroof_iron | 1.019 | 0.190 | 0.271 | 0.059 | -0.215 | -0.172 |
| Wroof_tile | 1.081 | 0.203 | 0.294 | 0.002 | -0.316 | -0.005 |
| Wroof_straw | 1.085 | 0.222 | 0.241 | 0.008 | -0.253 | -0.764 |
| Wmax_educ_none | 0.900 | 0.217 | 0.126 | 0.206 | 0.072 | 0.049 |
| Whead_female | 0.526 | 0.225 | 0.250 | 0.213 | -0.020 | -0.043 |
| Wmax_educ_fem_5to7 | 2.492 | 0.670 | 0.140 | 0.128 | -0.030 | -0.118 |
| Wtoilet_flushseptik | -0.376 | 0.130 | 0.302 | 0.010 | 0.110 | 0.049 |
| Wroof_mud | 0.760 | 0.243 | 0.051 | 0.007 | -0.034 | 0.005 |
| Wtoilet_none | -0.783 | 0.123 | 0.444 | 0.771 | -0.256 | -0.435 |
| Wwater_piped | 0.220 | 0.089 | 0.551 | 0.571 | 0.004 | -0.029 |
| Wowns_fridge | 1.981 | 0.579 | 0.027 | 0.000 | -0.054 | -0.017 |
| meanht | 0.194 | 0.061 | 1.169 | 3.126 | 0.380 | 0.288 |
| popdens | 0.000 | 0.000 | 588.223 | 11.950 | -0.022 | -0.004 |
| _cons | -3.500 | 0.473 | | | | |

Focusing first on the global influence diagnostics in column six, a value that is negative is contributing a lower average WHZ score in ilaka 901 compared to the rest of the country, which in turn is causing a higher wasting rate. The global influence diagnostics show that over 70% of the variables are negative, thus contributing to a lower average WHZ and so a higher wasting rate compared to the country as a whole. There were several variables that were having a larger contribution, such as less households in the area having brick walls or iron, tiled or straw roofs and more households being without a toilet. Also there are variables such as *meanht* that are having a positive effect on the WHZ and helping to lower the wasting rate. The *meanht* is positive as ilaka 901 is located in the mountains, hence is situated higher than the average small area. If we just focused on the global influence diagnostics of ilaka 901, there were a number of variables causing the small area to have a high wasting rate relative to the rest of the country, but evidently no one variable was causing the large wasting rate.

It is important to not only look at the influence diagnostics for the ilaka 901 in isolation, as it may be helpful to compare them to the influence statistics from all the other small areas. Figure 6.2 shows the global influence statistics for all the small areas for each of the variables. The figure shows the distribution of the influence statistics for each one small area for each variable. This results in there being 976 points for each variable (one for each small area). Ilaka 901 is shown as the black hollowed circle whereas the remaining 975 variables are each displayed by a grey dot. When comparing all the global influence diagnostics there are variables such as *girl* that has a very small spread of deletion diagnostics and nearly all sit on zero. This is because most small areas tend to have the same proportion of girls and boys, which seems to even out at about 0.5 when the population increases in each small area. The variable *terai* is a binary indicator, as a small area is either in the Terai (plains) or it is not, hence the influence statistic can take one of two values. To keep the focus on ilaka 901, a quick glance at Figure 6.2 suggests that none of the variables have particularly unusual influence statistics compared to the

Figure 6.2: Global deletion diagnostics for the small area wasting prevalence.

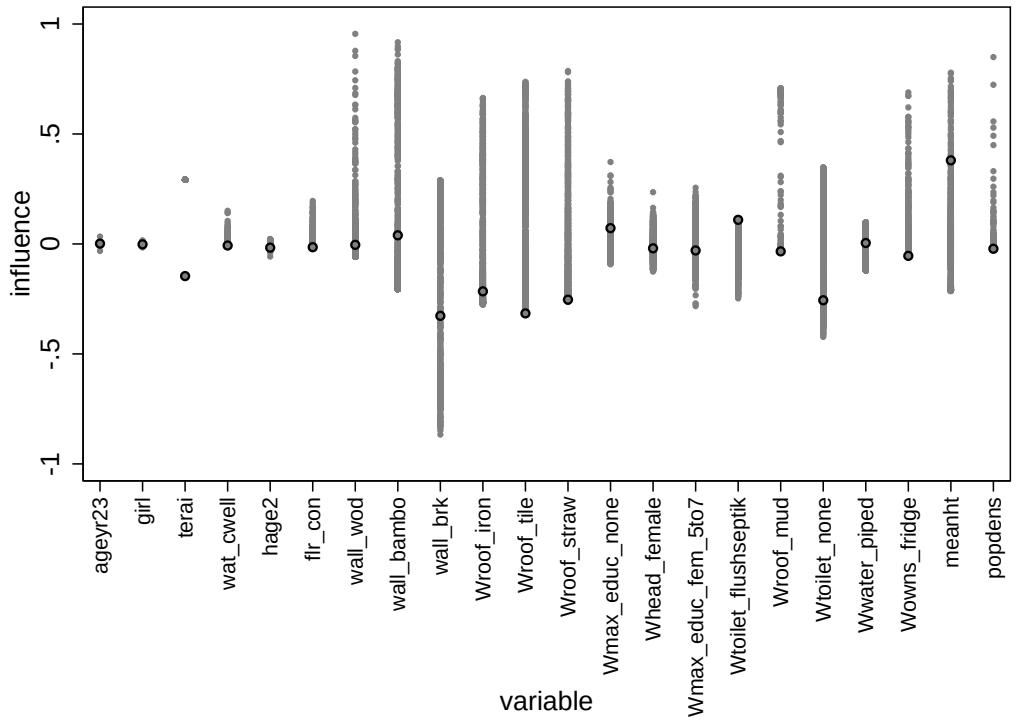
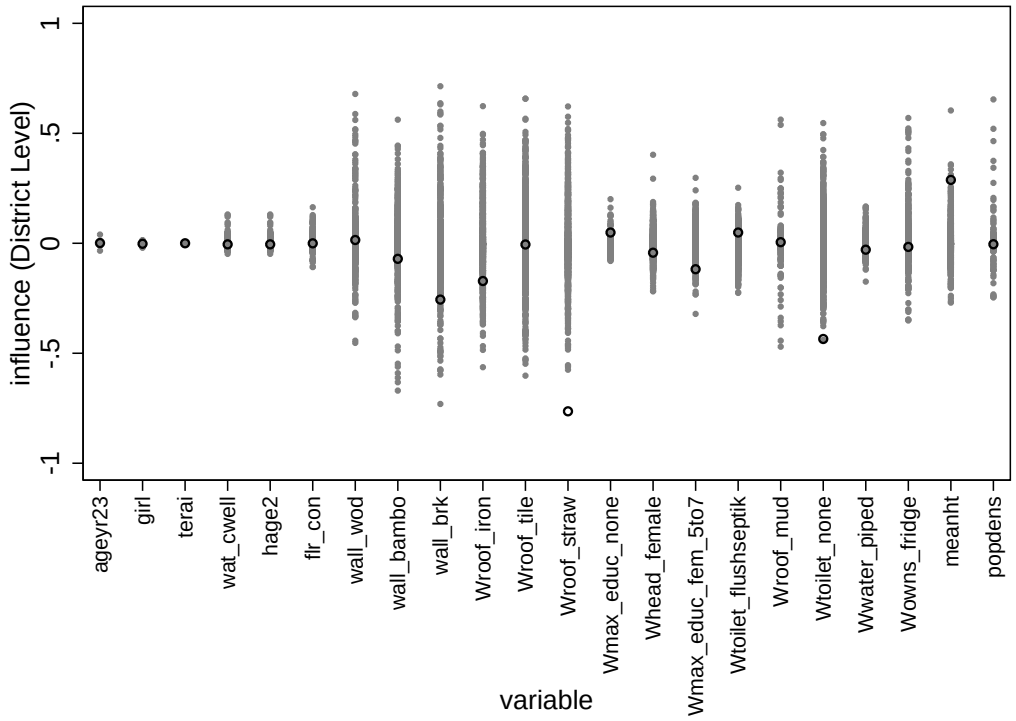



Figure 6.3: Localised deletion diagnostics for the small area wasting prevalence.



other small areas. This can be understood as none of the global influence statistics for ilaka 901 being largely different compared to the other small areas. Hence there did not appear to be one variable in particular that was causing this anomaly relative to the population as a whole.

In small area estimation, areas that are in close geographical proximity tend to be similar, however ilaka 901's estimated wasting rate was four times larger than the other ilakas in the district. Using (6.5) we can identify any discrepancy between ilaka 901 and the other ilakas in the district. Focusing on the localised deletion diagnostic in the final column of Table 6.1 we can discover which variable(s) are behaving differently in the small area compared to the remainder of the district. There are two variables behaving unusually, with one variable being more important than the other. The proportion of households in ilaka 901 without a toilet ($Wtoilet\_none$) is larger than many of the other small areas in the district as shown by the relatively large negative influence statistic of -0.435 (because the regression coefficient for this variable is negative and the influence diagnostic is negative, this means that the ilaka 901 has a higher proportion of people without a toilet than the other ilakas in the same district). However the greatest difference between ilaka 901 and the district mean concerns the variable $Wroof\_straw$, which is the proportion of households in a ward with a straw roof. Other ilakas in the district have a greater proportion of households with roofs made from straw than the overall population average, whereas ilaka 901 is less than average.

Figure 6.3 shows the distribution of the localised influence statistics for each small area. From the graph it is easy to see that the influence statistic for $Wroof\_straw$ is large and negative compared to the other statistics. This means that ilaka 901 is very different within its district for the proportion of houses with their roofs made out of straw, whereas the other small areas within their respective districts behave more similarly for this particular roofing type. Consequently either the households in ilaka 901 primarily use materials other that straw to construct their roofs while those in the ilakas nearby use straw as their primary construction material, or there

has been an error in the collection and processing of the census data.

Table 6.2 shows the breakdown of the ward level proportion of each of the roofing materials used for each ilaka in the Sankhuwasabha district; (where wards are the primary sampling units). In general the majority of roofs are straw, followed by iron. Ilaka 901 differed markedly, as less than 1% of the roofs are recorded as being straw, whereas most are classified as made out of "other material", coded as *other*. Having a high proportion of roof types classified as *other* is unusual as this proportion was less than 4% for 95% of all ilakas in Nepal. Furthermore after ilaka 901 the next largest ilaka in the country had 21.5% of the households recorded as having *other* roof types. This signals that the recorded 69% of households with *other* roofs for ilaka 901 could well be a mistake in the census data. The initial step of examining the coding of the variables showed that in the census *straw* was coded as 1 and *other* as 7. These are easily confused when handwritten as on the census forms, especially if 1 has been written with an initial upstroke. It seems possible then that when the raw census files were transferred onto the computer the 1's may have been mistakenly transcribed as 7's. This conclusion is however tentative, but fortunately there is an alternative to a special and expensive field visit. The Socio-economic database (Mega Publication and Research Centre, 2013) and a digital satellite imagery tool (Google Earth) were used to check the validity of the results for roof types. Both sources confirmed there were far more than 1% (and closer to 70%) of households with straw roofs in ilaka 901, indicating that a census miscoding error had occurred for this ilaka.

Rerunning the census is not feasible, so in order to correct the coding error for roof type, Bayesian multiple imputation was employed to reclassify the number of households with straw roofs incorrectly coded in ilaka 901. Denoting the number of households classified as *other* in PSU $j$ by $n_j$, the proportion misclassified as $\lambda_j$, and the number incorrectly classified as $Xstraw_j$, we have

Table 6.2: Proportion of households using each roofing material in the district of Sankhuwasabha.

| Ilaka | Straw | Iron | Tile | RCC | Planks | Mud | Other |
|---|---|---|---|---|---|---|---|
| 901 | 0.0074 | 0.0602 | 0.0024 | 0.0012 | 0.2255 | 0.0074 | 0.6959 |
| 902 | 0.5902 | 0.1797 | 0.0008 | 0.0040 | 0.0007 | 0.0019 | 0.2228 |
| 903 | 0.8773 | 0.0491 | 0.0031 | 0.0010 | 0.0029 | 0.0000 | 0.0666 |
| 904 | 0.8326 | 0.1553 | 0.0045 | 0.0007 | 0.0036 | 0.0000 | 0.0034 |
| 905 | 0.8265 | 0.1197 | 0.0018 | 0.0036 | 0.0027 | 0.0009 | 0.0447 |
| 906 | 0.4970 | 0.4632 | 0.0072 | 0.0161 | 0.0013 | 0.0001 | 0.0150 |
| 907 | 0.8868 | 0.0857 | 0.0081 | 0.0021 | 0.0020 | 0.0000 | 0.0154 |
| 908 | 0.7012 | 0.2615 | 0.0039 | 0.0020 | 0.0055 | 0.0000 | 0.0258 |
| 909 | 0.6461 | 0.3226 | 0.0173 | 0.0025 | 0.0022 | 0.0000 | 0.0092 |
| 910 | 0.8540 | 0.1301 | 0.0092 | 0.0029 | 0.0000 | 0.0000 | 0.0038 |
| 911 | 0.4638 | 0.4931 | 0.0158 | 0.0034 | 0.0005 | 0.0000 | 0.0234 |

$$Xstraw_j|\lambda_j \sim Binomial(n_j, \lambda_j) \tag{6.6}$$

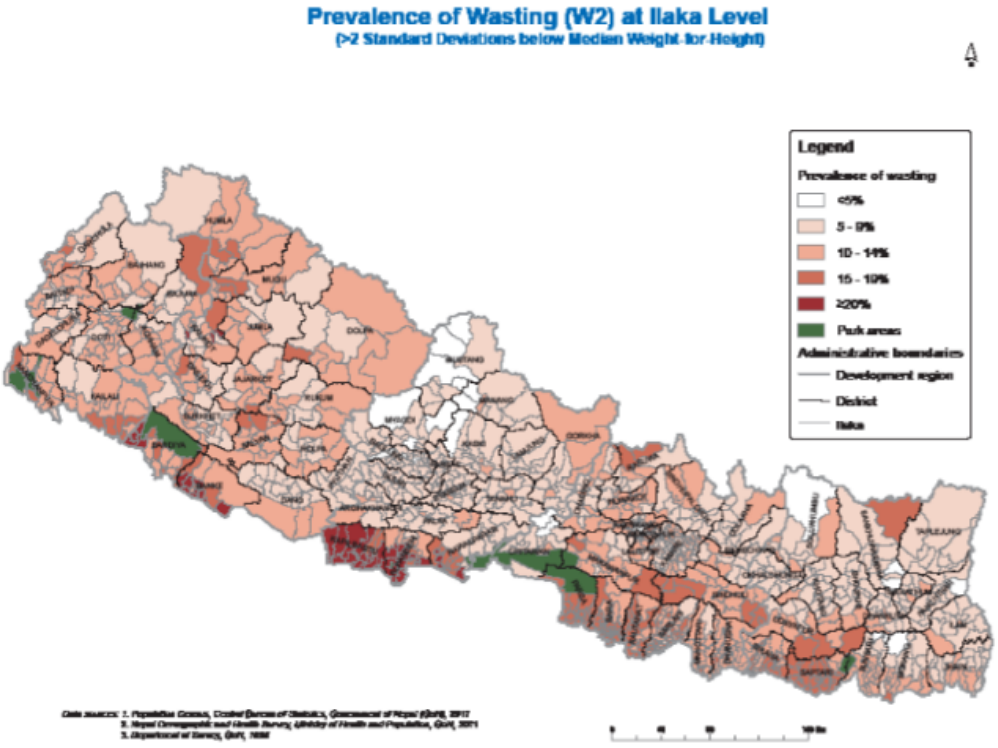
$$\lambda_j|\alpha_1, \alpha_2 \sim Beta(\alpha_1, \alpha_2) \tag{6.7}$$

where $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$, are parameters chosen to represent prior knowledge from UN World Food Programme experts about the $\lambda_j$s; these represent the best prediction of the proportion of households with straw houses and the uncertainty surrounding this. Multiple random draws of $\lambda_j$ and $Xstraw_j|\lambda_j$ give imputed values for the true $Wroof\_straw$ that are used to re-calculate the small area estimate.

Making changes to this one variable had major implications for the final estimated rate of wasting in ilaka 901, changing it from 44% (se 8.6%) to 19% (se 4.4%); this is shown in the corrected map in Figure 6.4. The revised wasting rate is still a cause for concern in terms of child welfare and the malnutrition, as anything over 5% is regarded as problematic; however for ilaka 901 the revised small area estimate suggested that the wasting problem in children under

five is far less severe than initially thought.

Figure 6.4: Corrected small area estimates of the prevalence of Wasting in Nepal.




### 6.3.1 Generalizing SAE Influence Diagnostics for the Nepalese Wasting Rate

Instead of focusing on one particular small area that has been identified as unusual and identifying the variables driving the peculiarity, the influence diagnostics can be used to identify *any* small area level means that are unusual. This could be done by a statistical rule based method or again by visual diagnostics. Figure 6.5 presents the SAE influence diagnostics for the variables included in the preliminary regression model fitted to WHZ for every ilaka in Nepal. In the preliminary model there are 22 variables fitted, shown in Table 6.1, resulting in $22 \times 976 = 21472$

Figure 6.5: Boxplot of global deletion diagnostics for the Wasting prevalence in Nepal.

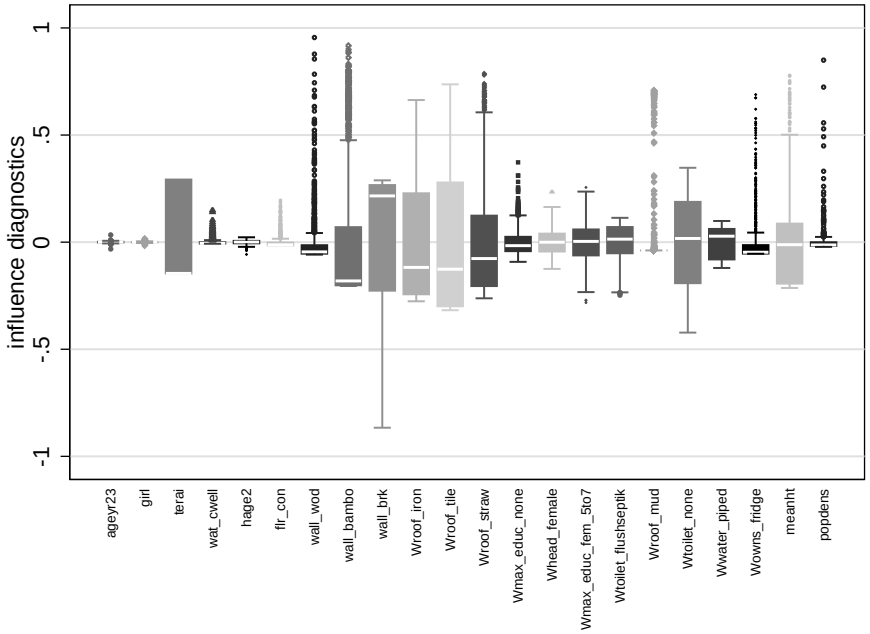
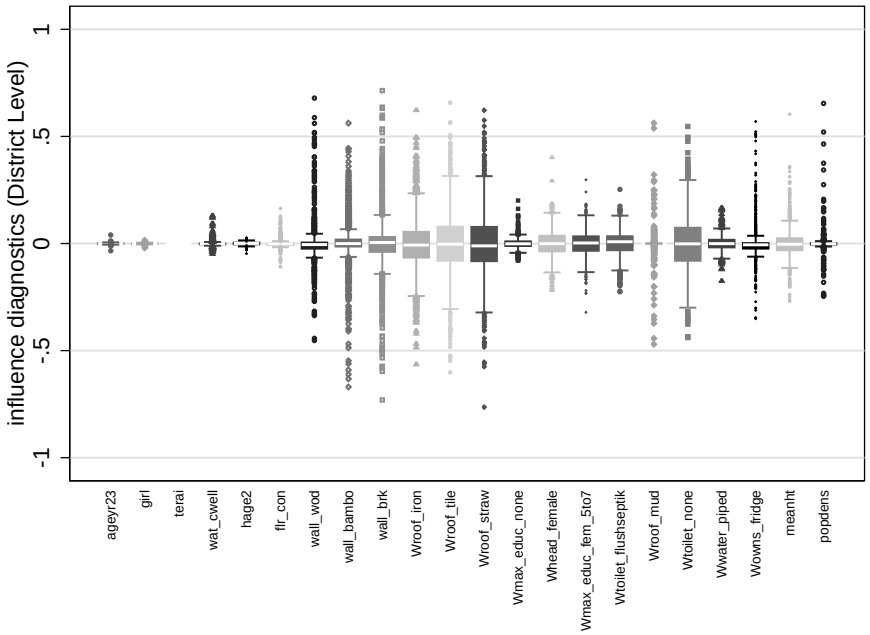



Figure 6.6: Boxplot of localised deletion diagnostics for the Wasting prevalence in Nepal.



influence statistics. Rather than using rule based methods, I used a series of boxplots to display the distribution of the influence statistics for each small area. Any points (small areas) that visually appeared to be different and fall relatively far away from the other influence statistics were further investigated. The diagnostic at this stage is based on what appears unusual visually, rather than a formal rule or numerical threshold, as a numerical threshold has the risk of excluding potentially influential statistics if it is placed too high and including too many data points if it is placed too low.

Figure 6.5, uses the global diagnostics ($\bar{x}_{ip} - \bar{\bar{x}}_{(i)p}$). This shows several values that stand out as being particularly unusual. Some of the variables involved are $wall\_wood$, $Wmax\_educ\_none$, $Wroof\_mud$, $Whead\_female$, $Wowns\_fridge$ and $popdens$. However further examination of these variables suggest that no obvious mistakes have been made but rather these small areas just remain relatively different compared to the rest of the country. For example the small area that appears to have a large positive influence diagnostic for $popdens$ is one of the ilakas in Kathmandu. Kathmandu however is very densely populated compared to the rest of the country, so this explains the difference.

Focusing on the locally centred diagnostics (6.5) shown in Figure 6.6, not only does $Wroof\_straw$ appear to be unusual for ilaka 901, but also there appears to be some other unusual small areas in several of the variables. One of these values is for $Wwater\_piped$, where in the district of Lalitpur there is one ilaka that has a smaller percentage of households with piped water compared to the remainder of the district. This can be shown in Table 6.3, where ilaka 2505 has only 5% of the population with piped water, whereas most of the other small areas in this district have over 80% of the households with piped water for drinking. Further investigation again revealed a large proportion classified as $other$, suggesting that another mistake may have been made when transferring raw data into the computer. Interestingly, the underlying coding issue seems to be similar to ilaka 901 with the coding for 1 being $piped$ and for 7 being

Table 6.3: Proportion of households using each type of drinking water in Lalitpur.

| District | Ilaka ID | Piped | Tube* [1] | Covered* | Uncovered* | Spout | River | Other |
|---|---|---|---|---|---|---|---|---|
| Lalitpur | 2501 | 0.82 | 0.00 | 0.04 | 0.01 | 0.05 | 0.00 | 0.07 |
| Lalitpur | 2502 | 0.82 | 0.00 | 0.01 | 0.02 | 0.14 | - | 0.01 |
| Lalitpur | 2503 | 0.90 | 0.00 | 0.03 | 0.02 | 0.04 | 0.00 | 0.00 |
| Lalitpur | 2504 | 0.81 | 0.00 | 0.06 | 0.02 | 0.07 | 0.02 | 0.02 |
| Lalitpur | 2505 | 0.05 | 0.01 | 0.38 | 0.04 | 0.00 | 0.00 | 0.51 |
| Lalitpur | 2506 | 0.93 | 0.00 | 0.02 | 0.00 | 0.01 | - | 0.00 |
| Lalitpur | 2507 | 0.48 | 0.06 | 0.28 | 0.05 | 0.01 | 0.00 | 0.07 |
| Lalitpur | 2508 | 0.85 | 0.00 | 0.04 | 0.04 | 0.05 | 0.00 | 0.00 |
| Lalitpur | 2509 | 0.91 | 0.00 | 0.01 | 0.01 | 0.06 | 0.00 | 0.00 |
| Lalitpur | 2510 | 0.91 | 0.00 | 0.04 | 0.01 | 0.02 | 0.01 | 0.00 |
| Lalitpur | 2511 | 0.96 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 |
| Lalitpur | 2512 | 0.97 | - | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| Lalitpur | 2513 | 0.95 | 0.00 | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 |
| Lalitpur | 2514 | 0.61 | 0.01 | 0.11 | 0.02 | 0.05 | 0.00 | 0.19 |

*other*.

Another small area that is behaving relatively differently with the variable $Wwater\_piped$ is ilaka 2704 within the district of Kathmandu. Table 6.4 shows that only about 8% of the small area have piped drinking water, whereas the rest of Kathmandu is somewhat higher. Instead there is a higher proportion of households that have their drinking water coming from a tube well in this small area. Further investigation would be needed to see if the drinking supply in this small area is really different or if this is a mistake. For both these small areas the lower proportion of piped drinking water leads to a lower predicted WHZ for each child and this in turn increases the wasting rate in the small area. The wasting rates in these two areas were 8.7% and 11.9% respectively. Neither of these estimates are particularly high compared to the remainder of the country, so even if there was a mistake it would not dramatically change the level of aid funding in the areas. Another example is for the variable $Wtoilet\_flushseptik$, where there appears to be a small area that has a relatively high influence of about 0.6. This would mean that the small area has a lower proportion of households using a septic tank than the remainder of

Table 6.4: Proportion of households using each type of drinking water in Kathmandu.

| District | Ilaka ID | Piped | Tube* | Covered* | Uncovered* | Spout | River | Other |
|---|---|---|---|---|---|---|---|---|
| Kathmandu | 2701 | 0.284 | 0.028 | 0.040 | 0.187 | 0.297 | 0.143 | 0.014 |
| Kathmandu | 2702 | 0.538 | 0.234 | 0.122 | 0.014 | 0.055 | 0.005 | 0.022 |
| Kathmandu | 2703 | 0.687 | 0.022 | 0.049 | 0.021 | 0.151 | 0.062 | 0.004 |
| Kathmandu | 2704 | 0.079 | 0.528 | 0.218 | 0.043 | 0.079 | 0.000 | 0.046 |
| Kathmandu | 2705 | 0.553 | 0.105 | 0.203 | 0.019 | 0.036 | 0.001 | 0.077 |
| Kathmandu | 2706 | 0.729 | 0.010 | 0.151 | 0.031 | 0.032 | 0.001 | 0.041 |
| Kathmandu | 2707 | 0.761 | 0.002 | 0.045 | 0.044 | 0.086 | 0.006 | 0.053 |
| Kathmandu | 2708 | 0.369 | 0.147 | 0.188 | 0.034 | 0.076 | 0.000 | 0.180 |
| Kathmandu | 2709 | 0.674 | 0.059 | 0.033 | 0.006 | 0.021 | 0.000 | 0.201 |
| Kathmandu | 2710 | 0.817 | 0.037 | 0.058 | 0.024 | 0.026 | 0.001 | 0.031 |
| Kathmandu | 2711 | 0.464 | 0.006 | 0.067 | 0.010 | 0.201 | 0.001 | 0.243 |
| Kathmandu | 2712 | 0.413 | 0.011 | 0.099 | 0.018 | 0.088 | 0.002 | 0.365 |
| Kathmandu | 2713 | 0.821 | 0.010 | 0.084 | 0.008 | 0.030 | 0.000 | 0.042 |
| Kathmandu | 2714 | 0.674 | 0.009 | 0.123 | 0.008 | 0.120 | 0.001 | 0.057 |
| Kathmandu | 2715 | 0.849 | 0.001 | 0.006 | 0.059 | 0.073 | 0.003 | 0.001 |
| Kathmandu | 2716 | 0.642 | 0.073 | 0.043 | 0.005 | 0.019 | 0.000 | 0.210 |
| Kathmandu | 2717 | 0.758 | 0.002 | 0.020 | 0.003 | 0.039 | 0.000 | 0.172 |

the district (as the product of a negative regression parameter and a negative difference results in a positive influence diagnostic). This influence diagnostic comes from ilaka 3501 in Chitawan district. Table 6.5 shows the proportion of households who use each of the four toilet types in this district. From the table it shows that ilaka 3501 has only 1% of the households using a septic tank, and instead the majority are using an ordinary toilet. This is a contrast to the remainder of the district, where the ilakas range from having 52% and 88% having a septic tank. Again further investigation would be needed to see if this is a true reflection, or if an error has been made. In general the overall small area estimated wasting rate of 4.6% is not overly different to the other small area wasting rates in the district, where these range from 3.3% to 9.1%. Therefore even if the variable was incorrect the change it would make would most likely not be too important. Figure 6.6 also showed the variables *wall_wood*, *wall_brk*, *wall_mud* and $Wtoilet\_none$ to be unusual, however further investigation into these variables didn't signal that

Table 6.5: Proportion of households using each type of toilet in Chitawan.

| District | Ilaka ID | Sewage | Septic tank | Ordinary | None |
|---|---|---|---|---|---|
| Chitawan | 3501 | 0.01 | 0.01 | 0.74 | 0.24 |
| Chitawan | 3502 | 0.01 | 0.67 | 0.23 | 0.08 |
| Chitawan | 3503 | 0.01 | 0.63 | 0.30 | 0.06 |
| Chitawan | 3504 | 0.01 | 0.66 | 0.25 | 0.07 |
| Chitawan | 3505 | 0.01 | 0.57 | 0.33 | 0.09 |
| Chitawan | 3506 | 0.03 | 0.52 | 0.33 | 0.12 |
| Chitawan | 3507 | 0.01 | 0.88 | 0.09 | 0.01 |
| Chitawan | 3508 | 0.01 | 0.86 | 0.10 | 0.03 |
| Chitawan | 3509 | 0.01 | 0.79 | 0.14 | 0.05 |
| Chitawan | 3510 | 0.02 | 0.72 | 0.21 | 0.06 |
| Chitawan | 3511 | 0.03 | 0.78 | 0.14 | 0.04 |
| Chitawan | 3512 | 0.00 | 0.73 | 0.22 | 0.05 |
| Chitawan | 3513 | 0.01 | 0.63 | 0.20 | 0.17 |
| Chitawan | 3514 | 0.01 | 0.71 | 0.22 | 0.05 |

they were anything to be concerned about.

## 6.4 Conclusion

In this chapter I have presented a method to quantify the relative importance of each variable in small area estimation using a new influence statistic. When a small area has been identified as unusual, the method provides a novel diagnostic tool for assessing which variable or variables are driving the irregularity. By considering the product of the regression coefficient and the difference between the population (or a localised) mean and the particular small area level mean for each variable, I have developed a diagnostic that indicates which variables are contributing to any anomalies in a given small area. The results suggest that using the locally centred means tends to be more sensitive. However it is important to check the anomalous results carefully, using local knowledge if possible, as many times these suspicious results can be explained.

The diagnostics can be presented graphically so that large contributions from each vari-

able for any small area can be detected visually. It might be thought useful to have a threshold for deciding whether a particular value of the influence diagnostic is large enough to cause concern. However Fox (1991) argues that numerical thresholds should be used with caution, and that graphical displays are more useful to assess which observations need further examination. Applying the SAE influence diagnostic to Nepal has demonstrated how the proposed influence statistic can identify which variables may be causing a small area to have an anomalous small area estimate. The significant change that was found in two particular small area estimates shows the importance of checking that both the model and the data are correct, as even one consistently miscoded auxiliary variable for a small area can lead to a large change in its SAE with possible serious consequences for aid allocation to the people living there.

# Chapter 7

# A Measure of Discriminatory Power for Small Area Estimates

In small area estimation the estimates need to be at an acceptable level of precision in order to be useful, for example the World Food Programme relies on precise estimates of food insecurity and child stunting and under-nutrition in order to target resources to small areas most in need. Therefore an important question is, how precise does the estimate need to be in order to be meaningful?

This chapter reviews some concepts and measures used to describe the precision in small area estimation, with a particular focus on the application to small area estimation of poverty. It proposes a measure of the overall precision that takes into account the uncertainty surrounding the estimates as well as the variability between the small areas. It also investigates how the ratio of the variability between the small area estimates to the uncertainty in the small area point estimates affects the level of precision needed to produce useful estimates.

## 7.1 Introduction

Precise estimates are important in order to convey reliable information. This is especially true in poverty mapping applications, where millions of dollars each year are distributed based on SAE in developing countries, as small areas with the highest level of deprivation get the highest amount of funding. In these situations the ranking is based on the point estimate in each small area. Therefore, it is very important the uncertainty surrounding the estimates is small, if the estimates are imprecise then the true ranking of the small areas in terms of poverty may not be a true reflection of the actual state of the country and funding won't be distributed effectively.

The precision surrounding the point estimates is usually measured in terms of the size of the standard errors associated with the small area point estimates. The standard error reflects the uncertainty surrounding the estimate. The true value is usually within plus or minus two standard errors of the estimate, hence the larger the standard error, the larger the uncertainty surrounding the estimate. The size of the standard error is dependent on a number of factors including the goodness of fit of the model fitted to the survey data and consequently the $R^2$ of the model (with adjustment for any random effects), the population size and the number of clusters in each small area.

In a variety of small area poverty applications the average standard error is usually required to be below 5% as otherwise the estimates would be regarded as too imprecise to be useful. An example for small area estimation for poverty is the ELL method and its extensions. In the original ELL paper (Elbers et al., 2003) it was recommended that the prediction standard errors should fall below 5% if the population in the small area has over 15,000 households. In subsequent applications the majority of the standard errors were kept below 5%. This can be seen in Bangladesh where the standard errors of the small area estimates varied from 0.3% to 11.5% with an average of 3.9% (Haslett and Jones, 2004). Another example is in Nepal where

the average standard error for the estimates was 3.5%, with approximately 15% of the small areas having a standard error above 5% (Haslett et al., 2014b). In Cambodia, the average standard error was 4.8% with the maximum being 14% (Haslett et al., 2013). This general rule can be seen in other poverty estimation examples such as in Ecuador where the standard errors for small areas in rural areas were averaging 6.7% (Haughton and Khandker, 2001). This was deemed too high as it created a 95% confidence interval where the true estimate could be within a 27% range, which is regarded as too imprecise for poverty estimation. The same principle is commonly used for small area estimation of other poverty indicators such as wasting, undernutrition and stunting rates. If the standard error is above 5% for an area, the population size of the small area can be increased, for example by combining two adjoining areas together, in order to reduce the corresponding standard error.

Juan-Albacea (2009) took a different approach and used the coefficient of variation (CV) to determine if the small area estimates were precise (where the CV is the standard error of the small area estimate divided by the poverty estimate). They suggest the rule of thumb that the CV needs to be below 10%.

Both these methods take into account the standard error, however what they fail to take into account is the variability between the small areas.

In the study of poverty estimation by Elbers et al. (2003), the inequality in Ecuador was decomposed into between and within group variability. It was found that at the small area level 85% of the variability in the rural areas could be attributed to the within group variability, whereas only 15% could be attributed to between group variability. This demonstrated that households within the small area were not homogeneous and the accuracy of the estimates should not solely be judged by the size of the standard error, as far more variability was occurring within the small areas rather than between the small areas. Targeting aid towards the areas with the highest poverty estimates is based on the assumption that households within the

small area are relatively homogeneous. The reliability of this assumption is tested in Tarozzi and Deaton (2009).

If the variability between the small areas is relatively small compared to the uncertainty of the estimates for the small areas, the estimates will not reliably be able to distinguish between different small areas. Let us take for example a poverty application. Let us assume the standard errors are relatively small, for example 5%. Let us also assume the variability between the small areas is also relatively small, for example the small area poverty rate ranges from 10%-25%. In this situation, the 95% confidence intervals reflecting the true small area statistics overlap. Therefore, it would be difficult to conclude which small areas are the most deprived. In poverty applications of small area estimation, it is imperative that the estimates are precise and reliable, as the estimates are often used to allocate millions of dollars of spending to the areas with the highest deprivation. A reliable ordering of small areas is very important as the small areas need to be precise enough to differentiate between them.

Not only is it beneficial to have a small standard error for each of the estimates, additionally having a relatively large standard deviation across the small area estimates will help to maintain the same ranking of the small area estimates as the 'true' values. This provides motivation to investigate not only the standard error of the small areas but also the variation between the small area statistics. I propose incorporating both these measures into a single statistic to measure the precision of the estimates by taking the ratio of the between small area standard deviation to the average standard error. An investigation into the relationship between this measure and the rank correlation is important, and the magnitude of these parameters needs to be considered as it is necessary to have a high rank correlation between the true small area poverty rate and the small area estimates. This will be explored in this chapter.

This chapter is organised into the following sections: Section 7.2 outlines the proposed methodology in the simple measurement error model framework, to analyse the level of pre-

cision needed in order to come up with meaningful estimates. A simulation study is applied in 7.2.1 to the simple measurement error model. Section 7.3 extends the analysis beyond the simple measurement error model, in order to explore if the relationship between the proposed measure and the reliability of the estimates holds in a complex data set typical of what would occur in poverty mapping applications. Conclusions are drawn in Section 7.4.

## 7.2 Methodology for the Simple Measurement Error Model

In this section, a measure is proposed that will help determine the level of precision SAE need to be in order to be useful. The method is outlined using the simple measurement error model (SMEM) and is based on the assumption that the properties of the SMEM hold. The proposed measure ($\kappa$) incorporates the variability between the small area estimates as well as the uncertainty surrounding these estimates. This measure will explore how changes in the relationship between the variability between the small area statistics and the error associated with them will affect the reliability of the estimates.

Suppose there are the true small area values $Y_i = 1, ..., I$, where $i$ denotes one of the $I$ small areas, where the $Y_i$ are either fixed unknown constants, or (in a model-based perspective) are random variables. The variance of the true values is denoted $\sigma_Y^2$; if the true small area values are regarded as fixed, this is defined as $\sum_{i=1}^{I}(Y_i - \bar{Y})^2/(I-1)$ where $\bar{Y}$ is the mean $\sum_{i=1}^{I} Y_i/I$. The $Y_i$ are unobservable; what is observed is the estimate of the small area statistic $\hat{Y}_i$, which also contains a measurement error $\varepsilon_i$.

Using the SMEM this can be defined as:

$$\varepsilon_i = \hat{Y}_i - Y_i \quad i = 1, .., I \tag{7.1}$$

where it is assumed that the $\varepsilon_i$ are independent and all have the same distribution, irrespective of the value of $Y_i$. We also assume that the estimation is unbiased, so each $\varepsilon_i$ has mean zero and variance $\sigma_\varepsilon^2$.

There are two main aspects affecting the precision required in small area estimation. The first aspect is the standard error of the estimates $\sigma_\varepsilon$, here assumed to be the same for all $Y_i$, and the second is the standard deviation of the true values $\sigma_Y$. However the true values are are unobservable, but can be measured indirectly by the standard deviation of the estimates $\sigma_{\hat{Y}}$:

$$SD(\hat{Y}) = \sigma_{\hat{Y}} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(\hat{Y}_i - \hat{\bar{Y}})^2} \qquad i = 1, ..., I \tag{7.2}$$

where $\hat{\bar{Y}}$ is the mean of the $I$ small-area estimates

There are several different measures of global imprecision that could have been used; one of these would be to take the average of the variances and then take the square root, furthermore the areas could have been weighted in terms of their population size. Rather the average standard error was selected here as this has commonly been used in previous poverty mapping applications. Additionally weighting was not used, as capital cities or larger cities tend be densely populated with low poverty rates and not much variation. If weighting was used these larger areas would be given more weight and the total imprecision would become lower. However the interest tends to be more on the smaller cities and the rural areas and the imprecision surrounding them, and if weighting is used, they are less prominent in the global measure of imprecision.

The statistic proposed to assess the level of precision required in SAE in order to generate reliable estimates is defined as

$$\kappa = \frac{SD(\hat{Y})}{SE(\hat{Y})} \quad = \quad \frac{\sigma_{\hat{Y}}}{\sigma_\varepsilon}. \tag{7.3}$$

This takes into consideration both the variation between the small areas as well as the standard error of the small area estimates.

In (7.1), $Y_i$ and $\varepsilon_i$ are assumed to be independent, so $V(\hat{Y}) = V(Y) + V(\varepsilon)$, therefore in this situation $\kappa$ can also be expressed as:

$$\kappa = \frac{\sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}}{\sigma_\varepsilon}. \tag{7.4}$$

If $\kappa$ was to increase it would reflect a greater variation in the SAE, a decrease in the average error associated with each estimate, or a combination of both. Alternatively, if $\kappa$ was to decrease it would reflect either a decrease in the variation of the SAE or an increase in the average error associated with each small area. Therefore the greater the level of $\kappa$ the easier it is to distinguish between the small areas. However, the value $\kappa$ needs to be above depends on the attributes that are important in SAE projects.

As previously mentioned there are a couple of important requirements in small area estimation. It is important that the small area estimates are a precise reflection of the true small area statistic; this can be assessed using the Pearson correlation, which measures the strength of the linear association between the true values and the estimates. In the SMEM (7.1) the Pearson correlation can be calculated as a function of the ratio of the between SAE variation to the standard error of the small area estimates ($\kappa$). As an illustration, this is shown below.

The Pearson correlation can be defined as:

$$\rho_{\hat{Y}Y} = \frac{E[\hat{Y}Y] - E[\hat{Y}]E[Y]}{\sigma_{\hat{Y}}\sigma_Y} \tag{7.5}$$

Substituting $\hat{Y}_i = Y_i + \varepsilon_i$ and $\sigma_{\hat{Y}}^2 = \sigma_Y^2 + \sigma_\varepsilon^2$ then

$$\rho_{\hat{Y}Y} = \frac{E[(Y+\varepsilon)Y] - E[Y+\varepsilon]E[Y]}{\sigma_Y \sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}} \tag{7.6}$$

$$\rho_{\hat{Y}Y} = \frac{E[Y^2] - E[\varepsilon Y] - (E[Y] - E[\varepsilon])E[Y]}{\sigma_Y \sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}} \tag{7.7}$$

$$\rho_{\hat{Y}Y} = \frac{E[Y^2] + E[\varepsilon]E[Y] - (E[Y] + E[\varepsilon])E[Y]}{\sigma_Y \sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}} \tag{7.8}$$

Assuming independence ($E[\varepsilon Y] = E[\varepsilon]E[Y]$), and since $E[\varepsilon] = 0$

$$\rho_{\hat{Y}Y} = \frac{E[Y^2] - E[Y]E[Y]}{\sigma_Y \sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}} \tag{7.9}$$

$$\rho_{\hat{Y}Y} = \frac{V[Y]}{\sigma_Y \sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}} \tag{7.10}$$

In a real world application we don't have the true values of the $Y$, therefore we replace

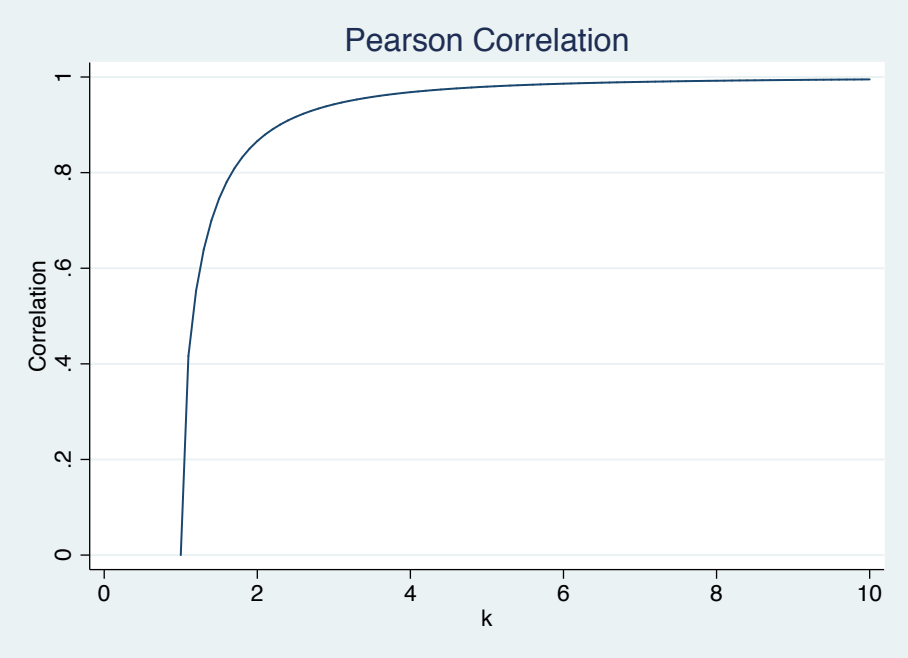
$$\sigma_Y^2 = \sigma_{\hat{Y}}^2 - \sigma_\varepsilon^2 \tag{7.11}$$

to give

$$\rho_{\hat{Y}Y} = \frac{\sigma_{\hat{Y}}^2 - \sigma_\varepsilon^2}{\sigma_{\hat{Y}} \sqrt{\sigma_{\hat{Y}}^2 - \sigma_\varepsilon^2}} \tag{7.12}$$

$$\rho_{\hat{Y}Y} = \frac{\sqrt{\sigma_{\hat{Y}}^2 - \sigma_\varepsilon^2}}{\sigma_{\hat{Y}}} \tag{7.13}$$

Figure 7.1: Pearson correlation as a function of $\kappa$.




$$\rho_{\hat{Y}Y} = \frac{\sqrt{\kappa^2 - 1}}{\kappa} \tag{7.14}$$

Figure 7.1 shows the Pearson correlation as a function of $\kappa$. It appears that when $\kappa$ is above three, the Pearson correlation is above 0.95 and hence the estimates will be relatively reliable. When $\kappa$ falls below two, there is a significant decrease in the reliability of the estimates. Figure 7.1 suggests that $\kappa$ might be useful for a simple diagnostics test of whether the small area estimates are sufficiently precise.

Another important requirement of SAE is that the ranking of the estimates should be similar to the ranking of the 'true' small area statistics, as the estimates needs to be reasonably accurate to effectively distribute funding to those most in need. The Spearman correlation can be used to measure this, since it is a non-parametric measure of the relationship between two variables that does not assume linearity. In this case, it is measuring the strength of the

relationship between the ranks of the true small area values ($Y_i$) and the estimates ($\hat{Y}_i$). Unlike the Pearson correlation, the Spearman rank correlation cannot be modelled explicitly as a function of $\kappa$, instead it is defined as

$$r_{\hat{Y}Y} = 1 - \frac{6\sum d_i^2}{I(I^2-1)} \tag{7.15}$$

where $d_i = rk(Y_i) - rk(\hat{Y}_i)$ is the difference in the small area's 'true' rank compared to the rank of its estimate and $I$ is the total number of small areas. Since the analysis is intractable for the rank correlation, we explore its relationship with $\kappa$ using simulation.

### 7.2.1 Simulation using the SMEM

This simulation is based on the model in (7.1) and is used to observe how changes to the ratio of between small area variation to the within small area error effects the reliability of the estimates. This will be observed by assessing how closely each of the small estimates matches its 'true' small area statistic, as assessed by the Pearson correlation. The simple simulation will also allow us to observe how the ranking of the small areas changes from their original ranking when $\kappa$ is adjusted.

The simulation study was set up by creating 1000 small area statistics $(Y_i)$. Each of these small area statistics was generated using a normal distribution, with a mean of 0 and standard deviation of 1, as can be seen in (7.16). These were defined to be the 'true' values. From here simulated values of $e_i$ were generated using (7.17) and added to the 'true' set of 1000 small area statistics generated from (7.16). This was repeated 100 times to give 100 sets of 1000 values each, and these were defined to be the small area estimates. The Pearson and Spearman correlation were then calculated for each set of the small area estimates, resulting in a total of 100 Pearson and Spearman correlation statistics.

$$Y_i \sim N(0,1) \qquad i = 1,...,1000 \tag{7.16}$$

$$\varepsilon_i \sim N(0,\sigma_\varepsilon) \tag{7.17}$$

After 100 simulations were run for each of the 1000 small areas for a particular value of $\sigma_\varepsilon$, the process was rerun adjusting the value of $\sigma_\varepsilon$, which consequently changed the value of $\kappa$. In total, the process was run for ten different values of $\sigma_\varepsilon$ in order to generate values of $\kappa$ from one to ten, or more precisely 1.01 to ten, as in the SMEM the values of $\kappa$ cannot be less than one. Via (7.4), $\kappa$ can only be one if all the small area values are equal.

The 100 simulations were used to compare the estimated values with the true values using the Pearson and the Spearman correlation. This was done to assess the reliability of each of the estimates for each simulation. For poverty mapping the ordering of the estimates is more important than the accuracy of the estimates, which makes the Spearman correlation a more useful tool.

Figure 7.2 and Figure 7.3 show the Pearson and the Spearman correlation respectively for each of the simulations. Each of the boxplots is based on 100 points, one for each of the simulations run. This is done for each value of $\kappa$ used.

The Pearson correlation and the Spearman correlation have similar results. When the ratio is below three both the Pearson and the Spearman correlation decrease noticeably, with an especially large decrease when $\kappa$ decreases from two to 1.01. The variation in the Pearson and Spearman correlation also increased as $\kappa$ decrease, which can be explained by having increased the standard error or a decrease in the variation between the small area estimates. There is little change in the correlations when $\kappa$ is between three and ten. When the ratio $\kappa$ is only 1.01, both

Figure 7.2: Pearson correlation of simulated data

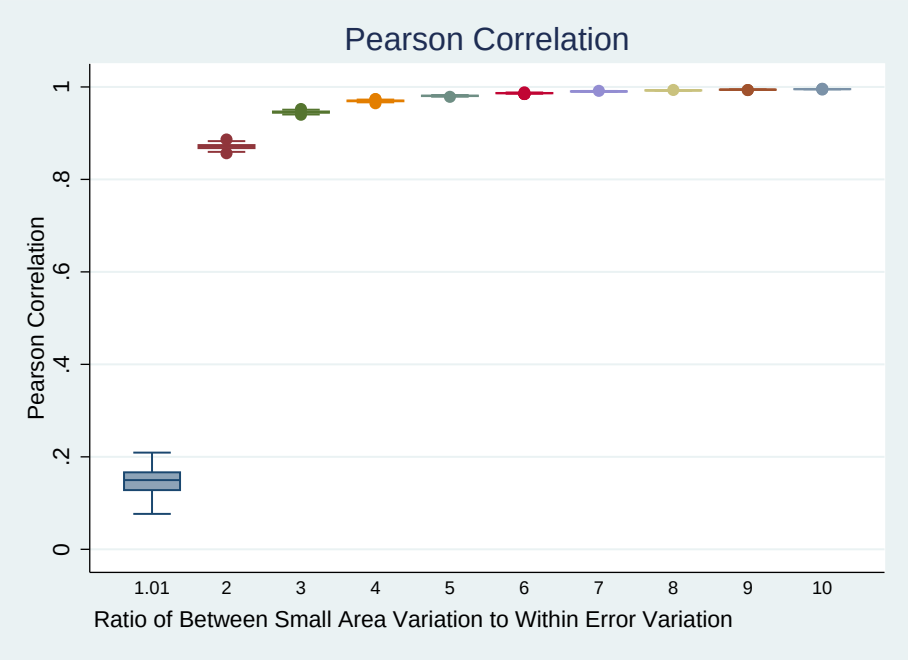
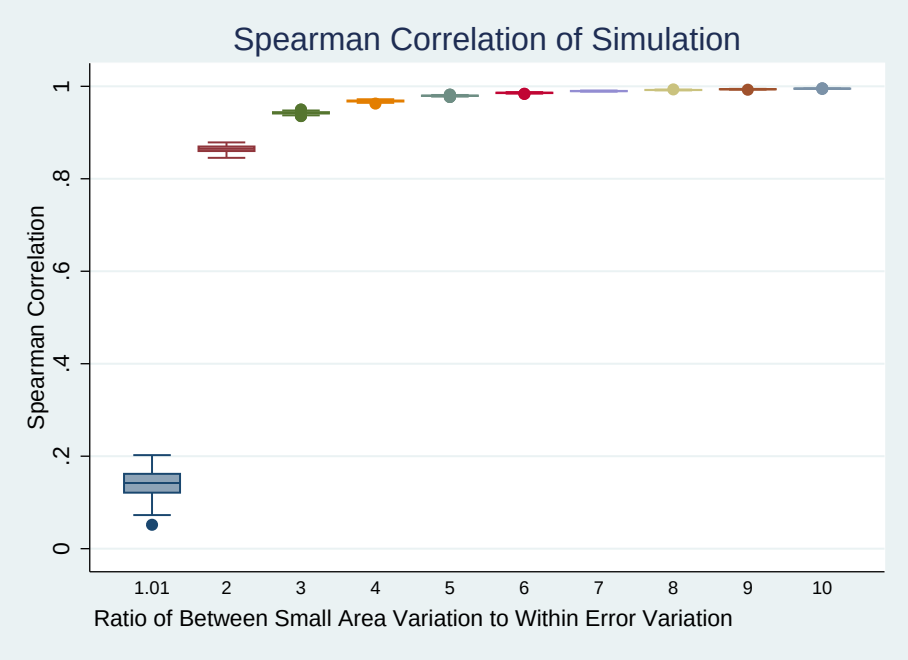



Figure 7.3: Spearman correlation of simulated data.



the Pearson and Spearman correlation are below 0.2 on average. When the ratio of $\kappa$ is above five the correlation between the simulations and the 'true' values is very high. In general if $\kappa$ is greater than three the ranking of the small areas, as measured by the Pearson correlation and the Spearman rank correlation, remains relatively accurate.

The simulated data gives an indication of how the ratio of the variation between the small areas compared to the error within the small areas affects the precision of the small area estimates. However, real data sets are more complicated compared to the SMEM, and therefore could lead to differing results. It is important to evaluate if the changes in $\kappa$ have the same effect on a more complex data set where the standard errors are not necessarily equal and may be related to the true values.

## 7.3 Beyond the Simple Measurement Error Model

In this section the measure is extended to an application beyond the SMEM. The previous section introduced the methodology on the SMEM in reality the situation is more complex and the assumptions of the SMEM such as independence and constant variance may not hold. Since the standard errors are no longer constant, $\sigma_e$ in the definition of $\kappa$ is replaced by the average standard error $ASE(\hat{Y}_i)$. This section applies the basic idea to a more realistic SAE application, where the assumptions of SMEM do not hold and algebraic analysis is intractable. This section will provide an indication if the relationship between the $\kappa$ and the reliability of estimates holds when applied a complex data set.

One particular form of small area estimation (ELL) is based on synthetic estimates generated from the unit level model (see Section 2.2). With this particular unit level model, a small area estimate $\hat{Y}_i$ can be defined as:

$$\hat{Y}_i = \frac{1}{N_i}\sum_{k=1}^{N_i}(\mathbf{X}_{ik}\hat{\beta} + \hat{u}_i) \tag{7.18}$$

where $N_i$ is the population in small area $i$ and $\hat{u}_i$ is the small area level variability. In many small area applications the 'small area' level variability ($u_i$) may not be modelled, rather the cluster level variability ($v_j$) as well as the unit level error is included as in (7.19), hence (7.18) can be adapted accordingly and the area level mean becomes

$$Y_{jk} = \mathbf{X}_{jk}\beta + \varepsilon_{jk} \quad where \quad \varepsilon_{jk} = v_j + e_{jk}. \tag{7.19}$$

$$\hat{Y}_i = \frac{1}{N_i}\sum_{k=1}^{N_i}(\mathbf{X}_{jk}\hat{\beta}). \tag{7.20}$$

Focusing on (7.20) there are several components affecting the ASE. Firstly, there is the uncertainty in $\hat{\beta}$, due to the model being fitted using survey data, or 'training data'. When the ELL method is used to generate the small area estimates, the model parameters are drawn from a multivariate normal distribution with mean $\hat{\beta}$ and $V(\hat{\beta})$. This source of uncertainty affecting the small area estimates is typically relatively low. The other two sources of uncertainty affecting the SAE are the cluster level uncertainty $v_j$, and the household level uncertainty $e_{jk}$, where their variability are denoted as $\sigma_v^2$ and $\sigma_e^2$ respectively. The more variation explained through the model at each particular level of aggregation, the lower the uncertainty and therefore error variance at that particular aggregation level. The standard error of a mean will decrease as the sample size increases. In general in SAE, the number of clusters sampled is much less than the number of households surveyed, meaning that the cluster level error will contribute more to the uncertainty in the estimates than the household level uncertainty. It then becomes important to ensure the unexplained cluster level variability is small (Haslett et al., 2014b). A large value of

$\sigma_v^2$ would indicate that there is a lot of unexplained variability within the cluster and therefore also within the small area and so wouldn't be as useful for SAE.

### 7.3.1 Poverty Simulation Study

This section uses a realistic data set in order to observe how changes in $\kappa$ affect the reliability of estimates. The simulation is based on the data set used to estimate the small area poverty rates in Cambodia (Haslett et al., 2013). In this simulation the standard error is adapted and this consequently affects $\kappa$ and the reliability of the estimates. In order to generate the small area estimates the survey data from the 2009 Cambodian Socio-Economic Survey (CSES) was used to generate simulated data and fit a model. The model was then used to make predictions for each of the 2,841,897 households included in the census, which were then amalgamated to give commune-level estimates. Full details of the original dataset are given in chapter 4.1. The following simulation observes how the changes in the value of $\kappa$ affect the reliability of the estimates when the data is transformed, skewed and not independent. This complex data set will help make generalisations about the relationship between the reliability of the estimates and the ratio $\kappa$.

In this small area estimation application, the poverty rate $(P)$ is the variable of primary interest. However it is not modelled directly, rather it is a non-linear transformation of expenditure $(Y)$. Furthermore, expenditure is highly right skewed so it is the log of expenditure that is modelled $(lnY)$. For further details refer to chapter 3.3.1 and chapter 4.1.3. Although the poverty rate is the outcome of interest, it will also be investigated how changes in $\kappa$ affect the small area estimation of log expenditure, shown in (7.21), and the expenditure (7.22) as well as the small area poverty rate in (7.23).

$$lnY_i^b = \frac{1}{N_i}\sum_{j,k\in i}^{N_i} \mathbf{x}'_{jk}\beta + v_j^b + e_{jk}^b \tag{7.21}$$

$$Y_i^b = \frac{1}{N_i}\sum_{j,k\in i}^{N_i} e^{(\mathbf{x}'_{jk}\beta + v_j^b + e_{jk}^b)} \tag{7.22}$$

$$P_i^b = \frac{1}{N_i}\sum_{j,k\in I}^{N_i} \left(\frac{z - Y_{jk}^b}{z}\right)^{\alpha} \mathbf{I}(Y_{jk}^b < z) \tag{7.23}$$

$j = 1,...,J_i \quad k = 1,...,N_{ij} \quad b = 1,...,100,$ following the notation of section 3.3.1

$\kappa$ cannot easily be changed directly, rather it must be changed through either making changes to the variation between the small area estimates, or by changing the average standard error associated with each of the small areas. In practice, it is easier to make changes to the average standard error, rather than trying to change the fitted model. In order to adapt the value of the average standard error for each of the small areas, the value of the cluster level variation as estimated in the original study, was multiplied by different values of an adjustment factor $c$ (this corresponds to $v_j$ being multiplied by different values in (7.21); this will be expanded on in the next paragraph.

In this application there are no census records for log expenditure, expenditure or the poverty rate. This means the true small area statistics are unknown, making it difficult to generate comparisons of the 'true' small area statistics to the small area estimate. Instead a statistic for each of the 1621 small areas was drawn from the superpopulation model (3.4) and these were defined as the 'true values'. Following this, 100 bootstrap estimates were simulated for

each small area using the ELL model outlined in chapter 3.3.1. These 100 bootstraps were used to measure the uncertainty surrounding each of the small area estimates. The small area estimates themselves were generated by taking the mean of the 100 bootstrap estimates for each of $lnY$, $Y$ and $P$ and the mean squared error was generated by taking the standard deviation across the 100 bootstrap estimates as can be seen in (7.24) to (7.26) respectively.

$$\widehat{lnY}_i = \frac{1}{B}\sum_{b=1}^{B} lnY_i^b \qquad MSE(lnY_i) = \frac{1}{B}\sum_{b=1}^{B}(lnY_i^b - \widehat{lnY}_i)^2 \tag{7.24}$$

$$\hat{Y}_i = \frac{1}{B}\sum_{b=1}^{B} Y_i^b \qquad MSE(Y_i) = \frac{1}{B}\sum_{b=1}^{B}(Y_i^b - \hat{Y}_i)^2 \tag{7.25}$$

$$\hat{P}_i = \frac{1}{B}\sum_{b=1}^{B} P_i^b \qquad MSE(P_i) = \frac{1}{B}\sum_{b=1}^{B}(P_i^b - \hat{P}_i)^2 \tag{7.26}$$

The mean and standard deviation are not usually produced for the log expenditure or the expenditure, as they are just used as steps in order to generate the small area poverty rate estimates. However, as well as investigating the sensitivity in $P_i$ to changes in $\kappa$, it was also investigated how the precision of the estimates of the normally distributed log expenditure $lnY$ as well as the highly right skewed distribution $Y$ were changed by changing the value $\kappa$. $\kappa$ was calculated by taking the ratio of the standard deviation of the means of the 100 bootstrap estimates to the average SE of the small area estimates as calculated from their respective bootstrap estimates. From here the 'true' values, that were generated from the superpopulation model, were compared to the small area estimates, using the Pearson correlation. Following this, the Spearman correlation was used to assess how well the ordering of the small area estimates reflects the ordering of the 'true' estimates.

The step by step process went as follows:

1. Use the superpopulation model from (3.4) to generate estimates of the log expenditure for each of the 1621 small areas. Define these to be the 'true' small area statistic. These are labelled as $lnY_i$, where as before $i = 1,..,1621$, denoting the small areas.

2. Use (7.21) to generate 100 bootstrap estimates for $lnY_i$. Each bootstrap draws a different estimate of $\beta$, $v_j$ and $e_{jk}$, as these values are not known. It draws $\hat{\beta}$ from the multivariate normal $(\hat{\beta}, V(\hat{\beta}))$ and draws the cluster and household level errors from an empirical distribution of $\hat{v}_j$ and $\hat{e}_{jk}$.

3. Take the mean and standard deviation of the 100 bootstrap estimates for each of the small areas using (7.24). These give respectively the small area estimate and its (estimated) standard error.

4. Gain the variability between the small areas by taking the standard deviation across the mean of the bootstrap estimates, shown as $SD(\hat{lnY}) = \sqrt{\frac{1}{I}\sum_1^I(\widehat{lnY_i} - \widehat{\overline{lnY}})}$, where $\widehat{\overline{lnY}}$ is the estimated mean log expenditure across all the small area estimates.

5. Gain the average measure of uncertainty of the small areas by taking the mean of the standard errors across the small areas $(ASE(\widehat{lnY}) = \frac{1}{I}\sum_1^I SE(\widehat{lnY}_i))$.

6. Using steps 4 and 5 generate the ratio $\kappa$.

7. Take the Pearson correlation of the 'true' small area statistics $Y_i$ (generated in step 1) with the small area estimates $\hat{Y}_i$ (where this is the mean of the 100 bootstrap estimates for each small area generated in step 2).

8. Calculate the Spearman correlation by comparing the rank of the 'true' small areas (from step 1) with the rank of the small area estimates (the estimates produced in step 2).

9. Follow steps 1-8 another 100 times. (Note that this can be done efficiently by swapping the "true value" generated in step 1 with one of the bootstrap values generated in step 2).

10. Repeat steps 1-9 for both the expenditure $(Y)$ and the poverty rate $(P)$.

11. Multiply the value of the cluster level error $v_j$ by the factor $c$, where c=0, 0.5, 1, 1.5, 2, 2.5 ,3 , 4 or 5 and repeat steps 1-10.

12. Plot the Pearson and Spearman correlation to compare the relationship between the ratio $\kappa$ and the correlation.

When changing $c$ (the factor that $\sigma_v$ was multiplied by), it was found that $\kappa$ for poverty estimation was much less sensitive to changes in $\sigma_v$ than the $\kappa$ for expenditure or log expenditure estimation. This is seen in Figures 7.4-7.9 where the values of $\kappa$ range from approximately 0.7 to 6.8 for poverty, whereas the values of $\kappa$ for log expenditure go from approximately one to 14 and from one to seventeen for expenditure. It is logical that $\kappa$ is less affected by changes in the cluster error, as the poverty rate is a non-linear function of log expenditure, which makes it less sensitive to changes in the cluster level error. Note that for the SMEM $\kappa$ can't be less than one, but in the presence of complex data sets it is possible for $\kappa$ to become less than one, as happened in the Poverty response since the assumptions of the SMEM no longer hold.

Figures 7.4 -7.9 show the relationship between $\kappa$ and the correlation for log expenditure, expenditure and poverty in Cambodia. This is also supplemented by Tables 7.1 to 7.3, where this illustrates the mean Pearson correlation and mean Spearman rank correlation for each of simulated value of $\kappa$. Furthermore, it shows the minimum and maximum Pearson and Spearman rank correlation for each value of $\kappa$.

The Pearson correlation is useful in order to compare how well the 1621 'true' small area statistics compare to the generated small area estimates. Supplementary to this is the Spearman rank correlation which assess how well the ranking of small area estimates compares to the 'true' small area statistics. In poverty applications of small area estimation, the Spearman rank correlation is of greater importance, as the order of the small areas in terms of their depths of

Figure 7.4: Pearson correlation for the log expenditure in Cambodia.

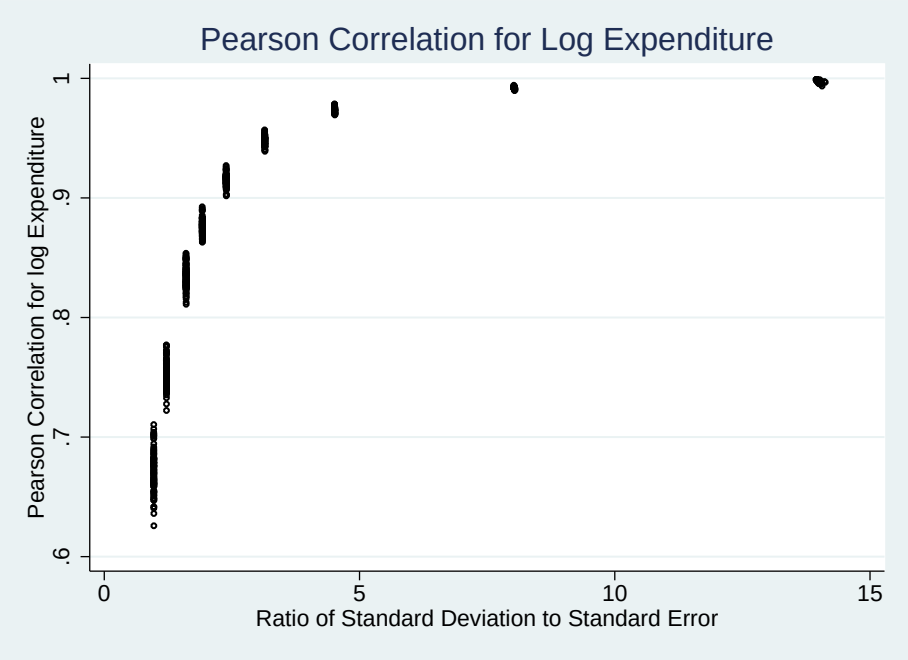



Figure 7.5: Spearman correlation for the log expenditure in Cambodia.

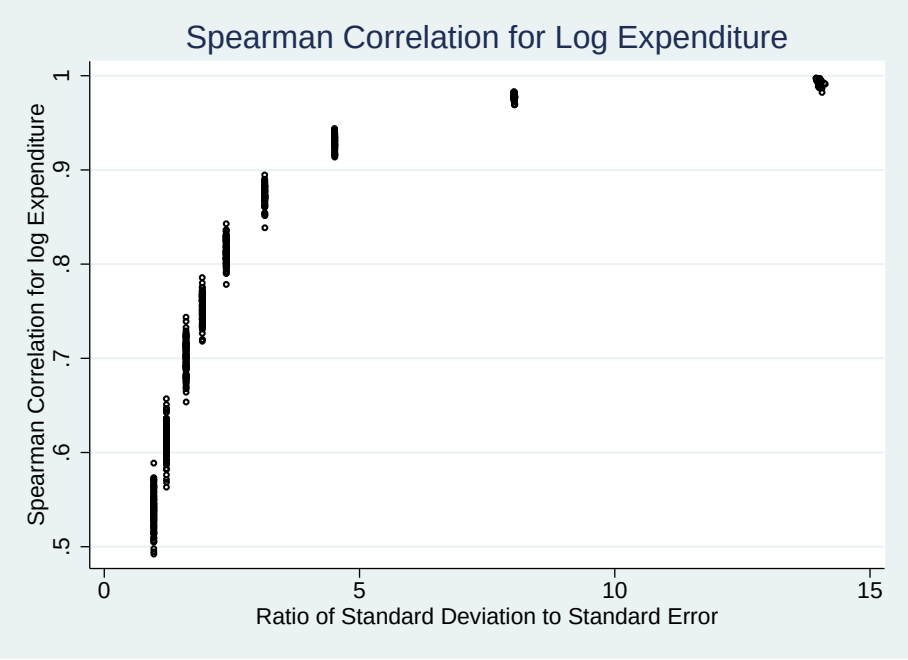


Figure 7.6: Pearson correlation for the expenditure in Cambodia.

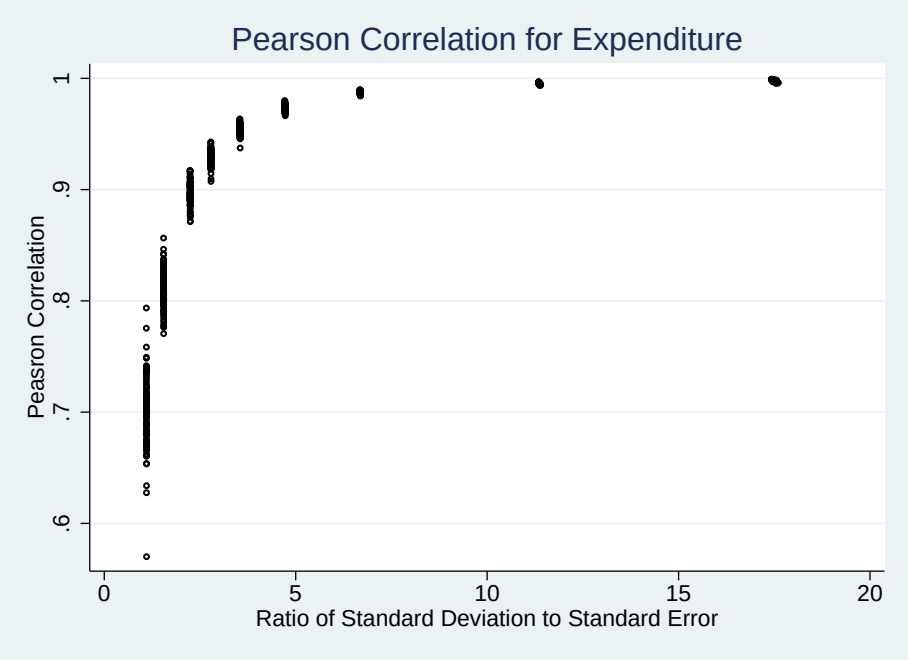



Figure 7.7: Spearman correlation for the expenditure in Cambodia.

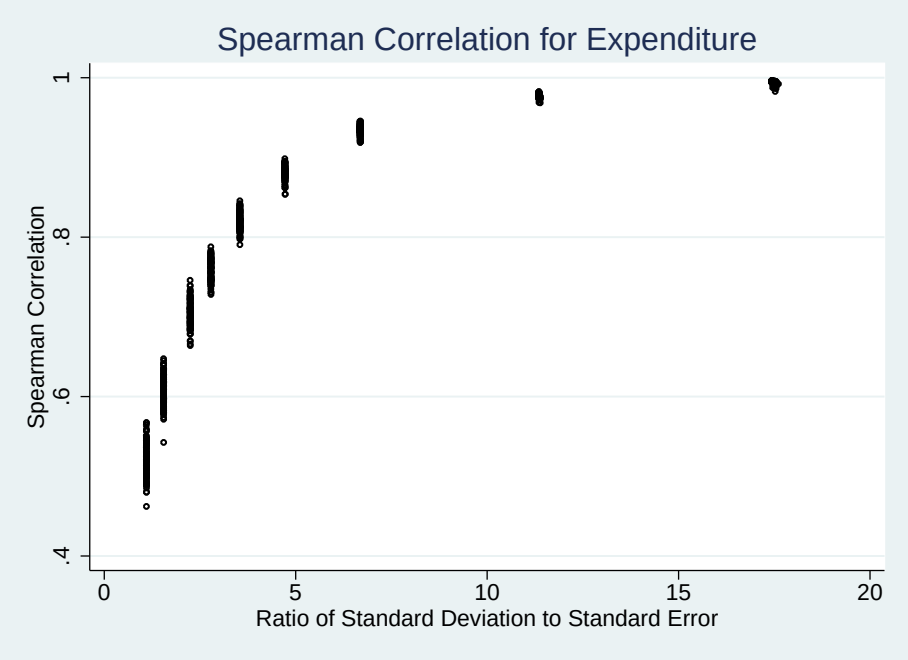


Figure 7.8: Pearson correlation for the poverty rate in Cambodia.

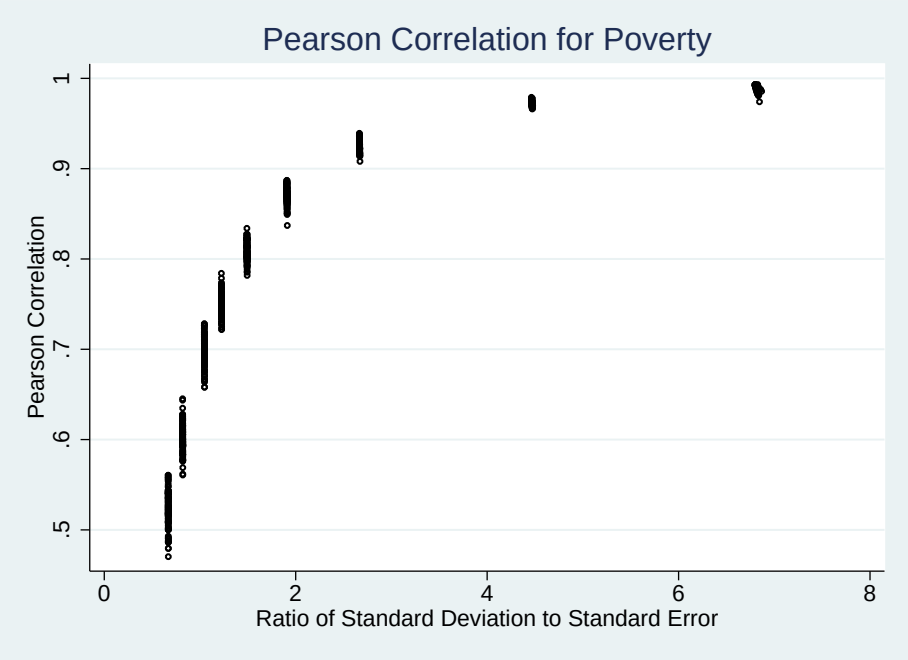
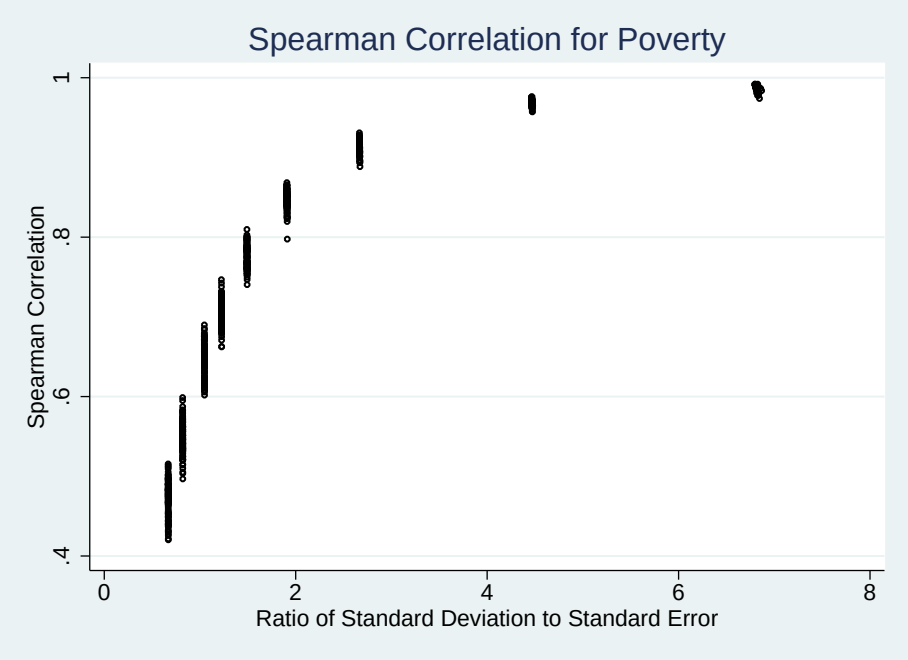



Figure 7.9: Spearman correlation for the poverty rate in Cambodia.



poverty is often used to determine which areas get funding.

In general all the figures show that as the ratio of the standard deviation to the standard error ($\kappa$) decreases, the range of correlation values increases. When $\kappa$ is high, the correlation is high and the variability of the correlation is also small, for example in Figure 7.6 when $\kappa$ is about 14, the Pearson and Spearman rank correlation are both concentrated around 0.99-1 for all 101 simulations. However when $\kappa$, decreases the Spearman rank and Pearson correlation decrease at an increasing rate. This is especially true when $\kappa$ is below five. Furthermore the lower the value $\kappa$, the greater the variation in the values of the Pearson and Spearman rank correlation, for example the Spearman rank correlation for expenditure ranges between 0.462 and 0.568 when $\kappa$ is approximately one.

Figures 7.4 and 7.5 show the Pearson and Spearman correlation for the log expenditure (where the log expenditure is normally distributed). It shows that when $\kappa$ is above five both the Pearson and Spearman correlation are above approximately 0.95, indicating that the estimates are relatively reliable. When $\kappa$ drops to approximately three the Pearson correlation is approximately 0.95, indicating that the values of the estimates still reliably represent the values of the 'true' small area values. Conversely, the Spearman rank correlation decreases to approximately 0.875, this means the order of the small areas in terms of their depth of poverty are not entirely reliable. When the ratio is approximately 1, the Spearman rank and the Pearson correlation decreases to an average of 0.54 and 0.67 respectively. This indicates at this lower value of $\kappa$ the estimates are not providing a reliable representation of the true small area statistics.

The expenditure plot has similar results as the log expenditure. When $\kappa$ is above three, the Pearson correlation is above 0.95. However, the Spearman correlation is less than 0.9 if $\kappa$ is under five. This would indicate for this particular distribution, the ratio $\kappa$ may have to be above six to seven in order to provide reliable results. This is a likely reflection of the highly right skewed distribution. Table 7.2 shows that when $\kappa$ is 1.101, the different simulations differ

Table 7.1: The effect $\kappa$ has on the Spearman and Pearson correlation for the log expenditure ($lnY_i$).

| | Spearman Correlation | | | Pearson Correlation | | |
|---|---|---|---|---|---|---|
| $\kappa$ | Mean | Min | Max | Mean | Min | Max |
| 13.985 | 0.995 | 0.982 | 0.997 | 0.998 | 0.994 | 0.999 |
| 8.030 | 0.978 | 0.969 | 0.983 | 0.992 | 0.990 | 0.994 |
| 4.513 | 0.931 | 0.914 | 0.944 | 0.974 | 0.969 | 0.979 |
| 3.144 | 0.875 | 0.839 | 0.895 | 0.949 | 0.939 | 0.957 |
| 2.389 | 0.813 | 0.778 | 0.843 | 0.915 | 0.902 | 0.927 |
| 1.920 | 0.753 | 0.718 | 0.786 | 0.876 | 0.863 | 0.893 |
| 1.603 | 0.701 | 0.654 | 0.744 | 0.834 | 0.811 | 0.854 |
| 1.217 | 0.615 | 0.563 | 0.657 | 0.753 | 0.722 | 0.777 |
| 0.970 | 0.540 | 0.492 | 0.589 | 0.672 | 0.626 | 0.710 |

largely in reliability, for example the Pearson correlation shows that for one of the simulations the correlation was only 0.57 indicating that it is not very reliable, whereas for one of the comparisons between $Y_i$ and $\hat{Y}_i$ the correlation was 0.794. In general, with the large range of correlations possible, it would be unwise to only have a ratio of $\kappa$ below 1.5 as the exact reliability would be difficult to determine.

In the Cambodian poverty application, the variable of interest is the poverty rate. Table 7.3 shows that when $\kappa$ is 2.67, the Spearman rank and Pearson correlation are 0.914 and 0.927 respectively. From Figure 7.8 and 7.9, it could be inferred that when $\kappa$ is above three, the Pearson and Spearman correlation would be above approximately 0.95. Therefore the small area estimates would be useful in predicting the small area statistics when $\kappa$ is above three and at an acceptable level of precision to effectively distribute aid to the areas most in need.

## 7.4 Conclusion

There is a strong relationship between the measure $\kappa$ and both the Spearman and Pearson correlation for log expenditure, expenditure and poverty. This can be shown explicitly for a simple

Table 7.2: The effect $\kappa$ has on the Spearman and Pearson correlation for the expenditure ($Y_i$).

| | Spearman Correlation | | | Pearson Correlation | | |
|---|---|---|---|---|---|---|
| $\kappa$ | Mean | Min | Max | Mean | Min | Max |
| 17.474 | 0.994 | 0.983 | 0.996 | 0.998 | 0.996 | 0.999 |
| 11.358 | 0.977 | 0.969 | 0.983 | 0.996 | 0.994 | 0.997 |
| 6.683 | 0.934 | 0.919 | 0.946 | 0.987 | 0.984 | 0.990 |
| 4.719 | 0.881 | 0.854 | 0.898 | 0.974 | 0.966 | 0.980 |
| 3.543 | 0.820 | 0.791 | 0.846 | 0.955 | 0.937 | 0.964 |
| 2.786 | 0.760 | 0.728 | 0.788 | 0.929 | 0.907 | 0.943 |
| 2.244 | 0.704 | 0.664 | 0.746 | 0.896 | 0.871 | 0.917 |
| 1.550 | 0.608 | 0.543 | 0.648 | 0.812 | 0.771 | 0.856 |
| 1.101 | 0.521 | 0.462 | 0.568 | 0.699 | 0.570 | 0.794 |

Table 7.3: The effect $\kappa$ has on the Spearman and Pearson correlation for the poverty ($P_i$).

| | Spearman Correlation | | | Pearson Correlation | | |
|---|---|---|---|---|---|---|
| $\kappa$ | Mean | Min | Max | Mean | Min | Max |
| 6.813 | 0.988 | 0.974 | 0.992 | 0.990 | 0.974 | 0.993 |
| 4.467 | 0.969 | 0.957 | 0.976 | 0.973 | 0.966 | 0.979 |
| 2.667 | 0.914 | 0.888 | 0.931 | 0.927 | 0.908 | 0.939 |
| 1.909 | 0.847 | 0.798 | 0.868 | 0.871 | 0.837 | 0.887 |
| 1.491 | 0.775 | 0.740 | 0.810 | 0.810 | 0.782 | 0.834 |
| 1.225 | 0.707 | 0.662 | 0.747 | 0.750 | 0.722 | 0.784 |
| 1.046 | 0.647 | 0.602 | 0.690 | 0.696 | 0.658 | 0.728 |
| 0.819 | 0.550 | 0.497 | 0.599 | 0.603 | 0.561 | 0.645 |
| 0.668 | 0.472 | 0.420 | 0.515 | 0.522 | 0.470 | 0.561 |

measurement error model where the Pearson correlation is a function of $\kappa$ and decreases as $\kappa$ decreases. The correlation is especially sensitive to changes in $\kappa$ when $\kappa$ is between one and two, where there is a vast decline in the correlation, however for all the three distributions the reliability of the estimates is not large enough to be useful. Furthermore, both the simple simulation and the more complex simulation show that the correlation is strongly correlated to changes in $\kappa$ where both the Pearson correlation and the Spearman correlation decrease when $\kappa$ decreases. In all three distributions, the correlations decrease at faster rate with lower values of $\kappa$.

Changes in the value of $\kappa$ only slightly affected the reliability of the estimates when $\kappa$ was above three for the Pearson correlation and five for the Spearman correlation. This was illustrated that in all three data sets (log expenditure, which is normally distributed, expenditure, which is highly right skewed and poverty which is a non-linear function of expenditure) there were only minimal improvement in the Pearson and Spearman correlation when $\kappa$ was above three and five respectively. Conversely, the decrease in the reliability of the estimates is a lot more prominent when $\kappa$ is less than three, in particular both the reliability in the SAE values as well as their ranking is very sensitive for expenditure. In general, if the focus was on how closely the estimates correlate to the 'true' value then as a rule of thumb the ratio $\kappa$ would need to be three or above. This seems to be particularly true for the poverty rate mapping examples, and holds irrespective of the distribution of the data. On the other hand if the focus was on ensuring the ranking of the small area estimates matched the ranking of the true small area statistics, then the value of $\kappa$ would be dependent on the distribution of the data. In the Cambodian poverty mapping example, the ratio $\kappa$ would need to be above five for the log expenditure (normal distribution) and log expenditure (right skewed distribution) in order to produce reliable estimates. However focusing on the variable of interest, the poverty rate ($P_i$), the ratio $\kappa$ would need to be three or more in order to produce a ranking of small areas that are

reliable estimates of the true small area statistics.

Although the average standard errors are traditionally used to determine if the estimates are precise, I propose that the ratio $\kappa$, which is the ratio of the standard deviation of the small area estimates to the average standard error of the small area estimates, is a better measure. This measure takes into account not only the uncertainty in the small area estimates but also the total variability across the small areas. Taking into account both these factors helps to more effectively predict the ability to distinguish between the small areas estimates and in general if the ratio remains above three, the rank order of estimates should be relatively stable and reliable. This however is a an empirical rather than a theoretical rule, therefore further examination using different datasets, models and small area estimation methods could be a further area of research.

# Chapter 8

# Conclusion and Future Work

This chapter will outline the results and conclusions that have been learnt throughout this research. It will also outline future directions to the research.

## 8.1 Conclusion

The aim for this research was to develop diagnostic techniques for small area estimation, primarily focusing on diagnostics for applications of SAE used for generating poverty rates. In the literature considerable attention has been given to developing various SAE techniques and adjusting existing methods to cater to particular applications, for example the Molina and Rao Empirical Bayes method (Molina and Rao, 2010), the Hierarchical Bayes method (Molina et al., 2014) and the ELL approach (Elbers et al., 2003) are all specialised approaches to estimate the poverty rate. However, much less attention has been given developing diagnostic measures to ensure the small area estimates are reliable. Millions of dollars of aid are distributed every year based on the predicted small area poverty estimates. Therefore, diagnostics become an important tool as it is essential the estimates are reliable, so that funding can be distributed efficiently

and effectively.

Existing diagnostics in SAE tend to focus on the model fitting phase, with a particular focus on how the 'training data' or the survey data affect the model parameters and reliability of the model. This thesis has rather focused on the diagnostics that relate to the reliability of the final small area estimates, therefore the diagnostics proposed incorporated not only the training data, but also the supplementary data used to make the predictions. In particular, there were three main diagnostics proposed:

- A variable importance metric to assess the importance of each of the variables included in the small area model.
- A deletion diagnostics to measure the influence a variable has on a small area estimate.
- A measure to quantify the level of precision needed in the small area estimates in order to generate reliable estimates.

Commonly, small area models in poverty estimation can be relatively complex, such as including over 30 variables. Although variables are providing predictive power at the model fitting phase, it does not necessarily equate to strength at the small area level. This is especially true for variables that are likely to have the same mean value in all small areas, or have a lot of variation within the small areas. Furthermore, when the models are large and complex there is the risk of spurious relationships occurring, which can result in small but spurious standard errors. Chapter Five proposed a variable importance metric (VIM) for small area estimation that incorporated not only the training or survey data used to fit the model, but also the supplementary data used to generate small area estimates. The VIM took the difference between the population mean and the small area mean for a particular variable and combined this with the variable's estimated regression coefficient.

This VIM was used to investigate methods for reducing model complexity while not significantly changing the estimated levels for each of the small areas. Commonly variables are judged as important in SAE based on their ability to explain the variation in the dependent variable (usually assessed via F-tests), and models are selected based on measures such as the adjusted $R^2$, whereas the proposed method also incorporates the distribution of the variable in the auxiliary data.

VIMs generated for the model were used for the poverty mapping exercise in Cambodia. Based on the VIMs, the least important variable was removed and the model was refitted. When reducing the complexity of the small area model it is important that the small area estimates remain accurate and precise. In poverty mapping exercises, it is also important the spatial distribution or the ranking of the small areas in terms of their depth of poverty remains the same. It was shown that in Cambodia the model could be reduced by seven to nine variables and still generate reliable estimates.

The VIM is useful in SAE as it provides insight into the variables that do not provide predictive power when estimates are aggregated up to small area level, and therefore may be able to be removed without affecting the estimates. Furthermore determining the variables that are the most important can be useful for the planning and collection of data in the future, as it would give an indication of the most important variables to collect, assuming the relationship between the variables remains the same. Although this was applied to the ELL method, it would have the similar results for methods such as the EBLUP, as it would apply the method to the synthetic component of the estimate, which is generally a lot larger than the direct component of the estimate, and therefore have similar results. In situations where the direct component of the EBLUP is large, this method may have differing results, which would lead to a future area of research, however in general the direct component is generally small, otherwise there would be no need for small area estimation.

Diagnostics to assess outliers and leverage are important tools in regression fitting activities. This is especially true when ensuring an observation or set of observations do not have undue influence on a model parameter. In SAE there have been methods proposed that are outlier robust, see for example Chambers and Tzavidis (2006). Although there has been previous research on how to identify and remedy outliers in the sampled data used to fit the model, there has been little research on identifying any influential or outlying observations in the auxiliary data that is used to generate the small area estimates. Chapter Six proposed a deletion diagnostics for small area estimation to identify any influential observations or variables in the auxiliary data that are having a large impact on the small area level estimate. The proposed deletion diagnostic incorporated the regression parameter for a particular variable with the difference between the small area level mean and the census mean, or a more localised mean.

This idea for the deletion diagnostics was motivated from a SAE exercise in Nepal, involving estimating the small area level wasting rate. In this application there was a particular small area that had a very high rate of children underweight for their age. Rather than using a rule based method to classify variables as unusual or not, a visual display was used to identify any patterns or unique observations in the census data. This method was able to identify that there had been an error in the collection and recording of the census data.

The proposed diagnostic is particularly useful in situations when a small area is shown to be unusual, as in the case of Nepal, as it can identify the variables that are having a large influence on the estimate. Alternatively the deletion diagnostic can be used as a general check after the small area estimates are generated to identify if there are any variables in any particular small areas behaving differently than expected. Performing this diagnostic after generating the estimates would be valuable in identifying any variables that are highly influential on a small area estimate, and possibly incorrect . This deletion diagnostic is important for policy makers to understand as it has the potential to avoid the ineffective distribution of money and resources.

Chapter Seven explored the level of precision the small area estimates need to be in order to be reliable. Precision of the estimates is an important factor needed in order to produce meaningful estimates. The reliability of small area estimates is usually based on the mean squared errors, or comparing the small area estimates with the direct estimates at some higher level of aggregation. If however the variation among the small area estimates is small, then the mean squared error would need to be proportionately smaller in order to be able to distinguish between the small area estimates, as once the uncertainty in the estimates is considered, the 95% confidence interval that the true means would fall between would likely overlap.

In small area estimation there are several different sources that contribute to the uncertainty of the estimates. In the model fitting stage the $R^2$, $\sigma_v^2$, $\sigma_e^2$ and the ratio $\sigma_v^2/(\sigma_v^2+\sigma_e^2)$ are measures of how well the model is fitted to the training data. These measures influence the standard error of the small area estimates. The lower the $R^2$ or the higher the value of the unexplained cluster variability the worse the fit of the model. In general these values as well as the standard errors of the estimates tend to be used as a gauge to the reliability of the small areas. This may not be an effective method to determine the reliability. Rather Chapter Seven proposed incorporating not only the standard error of the estimates, but also the variability between the small areas. The diagnostic measure $\kappa$ was proposed that takes the ratio of the standard deviation of the small area estimates to the average standard error of the estimates. It was found in general that if the standard deviation of the small area estimates is more than three to five times the size of the average standard error then the small area estimates will be reliable predictors of the rank order of the true small area statistics. However the distribution of the initial explanatory response in the training data influences the ratio needed for the small areas to maintain their true ranking. For the average expenditure, which is highly right skewed and the average log expenditure the ratio $\kappa$ would need to be above five in order to provide a reliable ranking ordering of the small areas. However $\kappa$ would only need to be above three in

order to provide a reliable ranking of the small areas in terms of the poverty rate. Therefore the value of $\kappa$ may depend not (just) on the distribution of the target variable at household level, but (also) on the area-level distribution of the statistic.

The ratio $\kappa$ is important as it considers not only the standard errors of the estimates, but also the variability in the small area estimates. It was demonstrated that the smaller the variation between the small areas the more precise the estimates would need to be and hence the smaller the standard error would need to be in order to produce reliable estimates.

Diagnostics for small area estimation are vital in order to ensure the reliability of the estimates. This is especially true in poverty mapping exercise when the estimates can be used to distribute millions of dollars of aid. Therefore it is essential for aid organisations and policy makers to understand the limitations of the estimates in order to prevent ineffective allocation of resources. Ideally field verification would be used to check the validity of the small area estimates. However this is costly and not always practical because of the geographic isolation of the small areas. Therefore, diagnostics of SAE become a useful tool to ensure that the estimates are reliable. Although, the diagnostics presented in this thesis are tailored to situations when there is unit level survey and census data, they can be adapted to situations when census data is unavailable and rather only survey data is available.

## 8.2 Recommendations and Future Directions

The usefulness of small area estimates depends on estimates being accurate and precise. This is where diagnostics becomes essential; to ensure the generated small area estimates are reliable. Diagnostics for small area estimation are largely unexplored. This thesis has just touched the surface. There are specific next steps that are possible, as well as more general future directions that could be considered.

In Chapter Five the small area estimates for the reduced models were compared to their respective small area estimates from the original model, with the assumption that they were independent of each other. This was used in (5.12) where there was no covariance term included when taking into account the standard error of $Pov_0 - Pov_r$. In reality there is covariance that exists from the estimation of model parameter $\hat{\beta}_0$, $\hat{\beta}_r$, however this would be expected to be a small contribution. A future step would involve finding a way to estimate the covariance term in the equation to avoid an underestimation of the $\zeta$.

In Chapter Six the proposed diagnostic compared the small area level means for the variables with the population means, or some more localised mean such as the district level mean. This diagnostic can take a localised mean at any level of aggregation for example at a regional level or ecological zone and these can be used to check if there are any outlying or influential areas in the particular region. Furthermore this diagnostic could be used to determine if there are clusters that are unusual rather than just small areas, as it is possible that there are specific clusters within a small area that are atypical. This would be achieved by taking the cluster level mean and subtracting the population level mean (or some localised mean). However this has the potential to create a large number of diagnostics to examine. It could be investigated if this would be a sensible thing to do, and if so what would the population size need to be in each cluster to generate reliable results. Furthermore a theoretical threshold value could be found for appraising this statistic.

Chapter Seven investigated the level of precision the estimates were required to be in order to be useful. However an important question remains: does the cluster and household level uncertainty evident in the training data reflect the true uncertainty that exists? The uncertainty applied to the small area estimates is a reflection of the cluster and household level uncertainty estimated by fitting a model to the training data. However the standard errors are conditional on the model being correct, and if the population isn't reflective of the survey data then the

small area estimates may be predicted at a higher or lower precision than is true. A diagnostic to assess the effect of model choice on the level of uncertainty would be useful to ensure the uncertainty surrounding the small area estimates is true. Furthermore one of the differences between the ELL method and EB_MR and the HB is that the ELL incorporates uncertainty at the cluster and household level, whereas the latter two incorporate small area level variation and household level variation. Therefore an extension would be to identify the level of precision the estimates would need if small area variation was recorded rather than cluster level variation.

More generally the diagnostics presented in this thesis focused on applications of SAE of poverty measures, where a model is fitted to the training data to explain a welfare measure. This is followed by a non-linear transformation and aggregation of the welfare measure to make predictions about the poverty measure. There are several methods that can be used to generate these poverty measures, where the availability and structure of the data as well as personal preference play a large role in determining which method is best to use. Although the diagnostics were suited to situations when unit record census and survey data were available they can be applied to applications that have survey data alone. A further step would be to ensure the diagnostics developed here are appropriate and can be used on small area estimates generated using other model based methods. For instances SAE methods such as the MR_EB, HB and the M-Quantile method have relatively larger small areas compared to the ELL. Each small area contains survey data, and a model is fitted to make predictions for the non-sampled data. This is then combined with the sampled data in each small area to make predictions. In these situations the proposed diagnostics can be adapted to the particular data structures and methods.

The M-Quantile method is useful to when there are outliers as it is free from distributional assumptions in the training data. However the deletion diagnostics proposed in Chapter Six would be useful to identify any non-sampled contributing to influential observations. In this situation the diagnostic would need to be adapted to adjust for how the model is fitted.

One of the main assumptions in SAE of poverty is that the training data or the survey data follows the same distribution as in the census. This tends to be checked at the country population level, but not so much at more disaggregated levels of the population. A diagnostic could be applied that compares the distribution of the variables of interest at a more localised level.

Statistical diagnostics are important in poverty estimation of SAE, however it is important to consult with local experts as they have detailed knowledge about the country. They will be able to provide invaluable insight into any particular data patterns and provide assistance in assessing the reliability of the estimates. Future work could look at how to incorporate expert opinion into the diagnostic process.

# Bibliography


Achen, C. H. (1982). *Interpreting and Using Regression*. Newbury Park, CA. Sage.

Arima, S., Datta, G. S., and Liseo, B. (2016). Models in Small Area Estimation when Covariates are Measured with Error. In Pratesi, M., editor, *Analysis of Poverty Data by Small Area Estimation*, chapter 8. John Wiley & Sons.

Baldermann, C., Salvati, N., and Schmid, T. (2018). Robust Small Area Estimation Under Spatial Non-Stationarity. *International Statistical Review*, 86(1):136–159.

Battese, G., Harter, R., and Fuller, W. (1988). An Error Components Model for Prediction of Country Crop Area Using Survey Satellite Data. *Journal American Statistics Association*, 9:28–36.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Observations and Sources of Collinearity*. John Wiley and Sons.

Breslow, N. (1996). Generalized Linear Models: Checking Assumptions and Strengthening Conclusions. *Statistica Applicata*, (8):23–41.

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001). Evaluation of Small Area Estimation Methods - An Application to Unemployment Estimates from the UK LFS.

Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2014). Outlier Robust Small Area Estimation. *Journal of Royal Statistics Society: Series B*, 76(1):47–69.

Chambers, R. and Tzavidis, N. (2006). M-Quantile Models for Small Area Estimation. *Biometrika*, 93(2):255–268.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall.

Coudouel, A., Hentschel, J., and Wodon, Q. (2002). *Poverty Measurement and Analysis. in PRSP Source book*. World Bank, Washington D.C., USA. http://siteresources.worldbank.org/INTPRS1/Resources/383606-1205334112622/5467_chap1.pdf.

Das, S. (2016). *Robust Inference in Poverty Mapping*. Doctoral thesis, University of Wollongong Australia.

Datta, G. S., Hall, P., and Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106(493):362–374.

Davis, B. (2003). *Choosing a Method for Poverty Mapping*. Poverty mapping. Food and Agriculture Organization of the United Nations.

Demidenko, E. and Stukel, T. A. (2004). Influence Analysis for Linear Mixed-Effect Models. *Statistics in Medicine*, 24:893–909.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1):pp. 355–364.

Elbers, C., Lanjouw, P., Lanjouw, J. O., and Bank., W. (2002). *Micro-level Estimation of Welfare*. World Bank, Development Research Group, Poverty Team Washington, D.C.

Fay, R. E. and Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James Stein Procedures to Census Data. *Journal of American Statistical Association*, 74:269–277.

Fletcher, D. (2009). A Brief History of the Khmer Rouge.

Foster, J., Greer, J., and Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica*, 52(3):761–766.

Fox, J. (1991). *Regression Diagnostics: An Introduction*. Sage Publications.

Ghosh, M. and Rao, J. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9(1):55–76.

Grömping, U. (2015). Variable Importance in Regression Models. *Wiley Interdisciplinary Review: Computational Statistics*, 7:137–152.

Hall, M. (2018). What Is Purchasing Power Parity (PPP)? . Investopedia. https://www.investopedia.com/updates/purchasing-power-parity-ppp/.

Haslett, S. (2012). Practical Guidelines for Design and Analysis of Sample Surveys for Small Area Estimation. *Journal of the Indian Society of Agriculture Statistics*, 66(1):203–212.

Haslett, S. (2016). Small Area Estimation using Both Survey and Census Unit Record Data: Links, Alternatives, and the Central Roles of Regression and Contextual Variables. In Pratesi, M., editor, *Analysis of Poverty Data by Small Area Estimation*, chapter 18. John Wiley & Sons.

Haslett, S. and Jones, G. (2004). Local Estimation of Poverty and Malnutrition in Bangladesh. Technical report, Bangladesh Bureau of Statistics and United Nations World Food Programme.

Haslett, S. and Jones, G. (2010). Small-Area Estimation of Poverty: The Aid and Industry Standard and its Alternatives. *Australian & New Zealand Journal of Statistics*, 52(4):341–362.

Haslett, S., Jones, G., Isidro, M., and Sefton, A. (2014a). Small Area Estimation of Food Insecurity and Undernutrition in Nepal. Technical report, Central Bureau of Statistics, National Planning Commissions Secretariat, World Food Programme, UNICEF and World Bank,, Katmandu, Nepal.

Haslett, S., Jones, G., Isidro, M., and Sefton, A. (2014b). Small Area Estimation of Food Insecurity and Undernutrition in Nepal. Technical report, Central Bureau of Statistics, National Planning Commissions Secretariat, World Food Programme, UNICEF and World Bank, Kathmandu, Nepal.

Haslett, S., Jones, G., and Sefton, A. (2013). Small Area Estimation of Poverty and Malnutrition in Cambodia. Technical report, National Institute of Statistics, Ministry of Planning, Royal Government of Cambodia and the United Nations World Food Programme.

Haslett, S., Noble, A., and Zababla, F. (2008). New Approaches to Small Area Estimation of Unemployment, The Official Statistics System, Wellington. *Official Statistics Research Series*, Vol 3.

Haughton, J. and Khandker, S. (2001). *HandBook on Poverty + Inequality*. The World Bank, Washington, DC. 978-0-8213-7614-0.

Henderson, C. (1953). Estimation of Variance and Covariance Components. *Biometrics*, 9(2):226–252.

Henninger, N. and Snel, M. (2002). Where Are The Poor ? Experiences with The Development and Use of Poverty Maps. World Resources Institute, Washington,DC.

Hentschel, J., Lanjouw, J. O., Lanjouw, P., and Poggi, J. (2000). Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador. *The World Bank Economic Review*, 14(1):pp. 147–165.

Ibp, I. (2011). *Cambodia Business Law Handbook Volume 1 Strategic and Practical Information*. International Business Publications USA.

International Development Committee (2010). *DFID's Programme in Nepal: Sixth Report of Session 2009-10, Vol. 1: Report, Together with Formal Minutes*. Number v. 1 in HC (Series) (Great Britain. Parliament. Great Britain: Parliament: House of Commons: International Development Committee: Stationery Office.

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *Ann. Statist.*, 36(4):1669–1692.

Juan-Albacea, Z. V. (2009). Small Area Estimation of Poverty Statistics. Technical report, Philippine Institute for Development Studies.

Kish, L. and Frankel, M. R. (1974). Inference from Complex Samples. *Journal of Royal Statistical Society. Series B (Methodological)*, 36(1):1–37.

Korn, E. L. and Graubard, B. I. (2003). Estimating Variance Components by Using Survey Data. *Journal of Royal Statistical Society. Series B*, 65(1):175–190.

Lahiri, P. and Suntornchost, J. (2015). Variable Selection for Linear Mixed Models with Applications in Small Area Estimation. *The Indian Journal of Statistics*, 77-B(2):312–320.

Lavrakas, P. (2008). Encyclopedia of Survey Research Methods.

Lee, E. S. and Forthofer, R. N. (2006). *Analyzing Complex Survey Data*. Quantitative Applications in the Social Sciences. Sage, second edition.

Lee, H. (2008). Jackknife variance estimation. In Lavrakas, P. J., editor, *Encyclopedia of Survey Research Methods*, pages 403–404. Sage Publications, Inc. doi = https://dx.doi.org/10.4135/9781412963947.n257.

Lehtonen, R. and Veijanez, A. (2009). *Handbook of Statistics Vol. 29B. Sample Surveys. Inference and Analysis*, chapter Design-Based Methods of Estimation for Domains and Small Areas, pages 219–249. Elsevier.

Li, J. (2007). *Regression Diagnostics for Complex Survey Data; Identification of Influential Observations*. PhD thesis, University of Maryland.

Li, J. and Valliant, R. (2009). Survey Weighted Hat Matrix and Leverage. *Survey Methodology*, 35(1):15–24. Statistics Canada.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press. ISBN: 0-534-35361-4.

Lohr, S. (2009). *Sampling: Design and Analysis*. Brooks/Cole CENGAGE Learning, second edition. 13-978-0-495-10527-5.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, 9(1):1–19. R package verson 2.2.

MacGibbon, B. and Tomberlin, T. J. (1989). Small area estimates of proportions via empirical bayes techniques. *Survey Methodology*, 15:237–252.

Malec, D. J. (2008). Superpopulation. In Lavrakas, P. J., editor, *Encyclopedia of Survey Research Methods*, pages 856–857. Sage Publications, Inc.

Marchetti, S., Tzavidis, N., and Pratesi, M. (2012). Non-parametric Bootstrap Mean Squared Error Estimation for M-Quantile Estimators of Small Area Averages, Quantiles and Poverty Indicators. *Computational Statistics and Data Analysis*, 56:2889–2902.

Mega Publication and Research Centre (2013). Village Development Committee and Demographic Profile of Nepal 2013: A Socio-economic Development Database of Nepal. Technical report. Kathmandu, Nepal.

Ministry of Health and Population (MOHP)[Nepal] (2012). Nepal Demographic and Health Survey 2011. Technical report, Kathmandu, Nepal: Ministry of Health and Population, New ERA, and ICF International, Calverton, Maryland.

Molina, I., Nandram, B., and Rao, J. N. K. (2014). Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach. *The Annals of Applied Statistics*, 8:852–885.

Molina, I. and Rao, J. N. K. (2010). Small Area Estimation of Poverty Indicators. *Canadian Journal of Statistics*, 38:369–385.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of Royal Statistics Society. Series A (General)*, 135(3):370–384.

Noble, A., Haslett, S., and Arnold, G. (2002). Small Area Estimation via Generalized Linear Models. *Journal of Official Statistics*, 18(1):45–60.

Pan, J., Fei, Y., and Foster, P. (2014). Case-Deletion Diagnostics for Linear Mixed Models. *Technometrics*, 3(56):269–281.

Pfeffermann, D. (2002). Small Area Estimation-New Developments and Directions. *International Statistical Review*, 70:125–143.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, 28(1):40–68.

Pfeffermann, D. and Sverchkov, M. (2009). *Handbook of Statistics 29B:Inference and Analysis*, volume 29B. Elsevier Oxford.

Preisser, J. S., Qaqish, B. F., and Perin, J. (2008). Miscellanea A Note on Deletion Diagnostics for Estimating Equations. *Biometrika*, 95(2):509–513.

Quenouille, M. H. (1949). Approximate Tests of Correlation in Time Series. *Journal of the Royal Statistical Society*, 11(B):68–84.

RamaRao, N. (2008). Background Paper on 2008 General Population Census of Cambodia. Paper, Cambodia National Institute of Statistics, Ministry of Planning.

Rao, C. R. (1972). Estimation of Variance and Covariance Components in Linear Models. *Journal of the American Statistical Association*, 67(337):pp. 112–115.

Rao, J. N. . K. and Molina, I. (2016). Empirical Bayes and Hierarchical Bayes Estimation of Poverty Measures for Small Areas. In Pratesi, M., editor, *Analysis of Poverty Data by Small Area Estimation*, chapter 18. John Wiley & Sons.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, New York.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation, 2nd Edition*. John Wiley & Sons Ltd, New York.

Rao, P. (1997). *Variance Components: Mixed Models, Methodologies and Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Saei, A. and Chambers, R. (2003). Small Area Estimation: A Review of Methods Based on the Application of Mixed Models.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag New York.

Searle, S., Casella, G., and McCulloch, C. (2009). *Variance Components*. John Wiley & Sons.

Shen, W. and Louis, T. A. (1998). Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 60(2):455–471.

Skinner, C. and Wakefield, J. (2017). Introduction to the Design and Analysis of Complex Survey Data. *Statistical Science*, 32(2):165–175.

Tarozzi, A. and Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *Review of Economics and Statistics*, 91(4):773–792.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29(2):614–614.

Tzavidis, N., Salvati, N., and Pratesi, M. (2008). M-Quantile Models with Application to Poverty Mapping. *Statistical Methods & Applications*, 17:393–411.

UN (1997). Human Development to Eradicate Poverty. Development Report, United Nations. url: http://hdr.undp.org/en/reports/global/hdr1997/.

UN Department of Economic and Social Affairs Statistics Division (2005). *Household Sample Surveys in Developing and Transition Countries*. United Nations, New York. ST/ESA/STAT/SER.F/96.

UNDP (2011). Cambodia, Country Profile: Human Development Indicators. **URL:** http://hdrstats.undp.org/en/countries/profiles/KHM.html.

UNICEF (2006). Situation of Children and Women in Nepal 2006. Technical report, Kathmandu, Nepal.

United Nations Development Programme (2014). Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience. Technical report, Human Development Report,, USA.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, 92(2):351–370.

Valliant, R. (2010). Linear Regression Diagnostics for Survey Data. In *Proceedings of the Statistical Society of Canada*, Quebec. Available at $http://www.ssc.ca/survey/SMSProceedings\_e.html$.

Van den Brakel, J. and Buelens, B. (2014). Covariate Selection for Small Area Estimation in Repeated Sample Surveys. *Statistics in Transition New Series and Survey Methodology*, 16(4):523–540.

Wang, Y. and Chen, H.-J. (2012). Use of Percentiles and Z-Scores in Anthropometry. In Preedy, V. R., editor, *Handbook of Anthropometry Physical Measures of Human Form in Health and Disease*, chapter 2. Springer.

WHO (2014). Children: Reducing Mortality. World Health Organization.

World Bank (2011a). How are Poverty Maps Built? **URL:**http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTPA/.

World Bank (2011b). Living Standards Measurement Study: Integrated Surveys on Agriculture.

World Bank (2015). FAQs:Global Poverty Line Update.

World Health Organization and UNICEF (2009). *WHO Child Growth Standards and the Identification of Severe Acute Malnutrition in Infants and*

*Children*. **URL:***http://www.who.int/nutrition/publications/severemalnutrition/ - 9789241598163_eng.pdf*.

Xu, R. (2003). Measuring Explained Variation in Linear Mixed Effects Models. *Statistics in Medicine*, 22(22):3527–3541.

# Appendices

# Appendix A

# Stata Code

```
*Chapter 5
*As a starting point there use the model and output formulated by Haslett et al (2013)

The original code by Haslett et al (2013) was adapted
Here this is re estimating the small area estimates after with the reduced models
clear
set memory 700m
set matsize 200

global sraw "E:\PhD\Cambodia_2017\Data\CSES2009\Raw"
global snew "E:\PhD\Cambodia_2017\Data\CSES2009\Created"
global craw "E:\PhD\Cambodia_2017\Data\Census2008\Raw"
global cnew "E:\PhD\Cambodia_2017\Data\Census2008\Created"
global panal "E:\PhD\Cambodia_2017\Analysis"
global results "E:\PhD\Cambodia_2017\Results"
global temp  "E:\PhD\Cambodia_2017\Data\Temp"
global part3 "E:\PhD\Cambodia_2017\Results\part3"

run "$panal\SURVReg_red.do"
run "$panal\SigEta.do"
run "$panal\SURVe2reg.do"

global j=1
while $j<=100 {
run "$panal\RanDraw_red.do"
global j=$j+1
}
```

```
\*here SigEta and SURVe2reg don't need to be changed-and can be seen later on in chapter 7s
code, so I won't list them here*/
*Where the various scripts from SURVReg_red.do are:

use $snew\CSES_povmodel.dta, clear
#delimit;
global xvar "hhsize lnhhsz pkids06 plit pseced notoilet numroom rfree
car cellphone computer electric motorbike phone radio tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 roof_cXS3 numroomXS3 motorbikeXS3";
#delimit cr

order $xvar
foreach var of varlist hhsize-motorbikeXS3{
sum `var'
replace `var'=(`var'-r(mean))/r(sd)
}

svy: reg ln_exp $xvar

*Final model (????) :
*Reduce:
#delimit;
global xvar "hhsize lnhhsz pkids06 plit pseced notoilet numroom rfree
car cellphone computer electric motorbike phone radio tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 roof_cXS3 numroomXS3 motorbikeXS3";
#delimit cr

svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing radio
*Reduce:
```

```
#delimit;
global xvar "hhsize lnhhsz pkids06 plit pseced notoilet numroom rfree
car cellphone computer electric motorbike phone tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 roof_cXS3 numroomXS3 motorbikeXS3";
#delimit cr
svy: reg ln_exp $xvar

* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio
*Reduce:
#delimit;
global xvar "hhsize lnhhsz pkids06 plit pseced notoilet numroom rfree
car cellphone computer electric motorbike tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 roof_cXS3 numroomXS3 motorbikeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3
*Reduce:
#delimit;
global xvar "hhsize lnhhsz pkids06 plit pseced notoilet numroom rfree
car cellphone computer electric motorbike tv
```

```
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3 motorbikeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom rfree
car cellphone computer electric motorbike tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3 motorbikeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone computer electric motorbike tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3 motorbikeXS3";
#delimit cr
```

```
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3 motorbikeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
```

```
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_t floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e resplus_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"
```

```
*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e reg3 tonlesap plnmount
hhsizeXS3 numroomXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_s roof_t roof_c roof_m wall_b
boat_e cellphone_e h_lit_e plit_e reg3 tonlesap plnmount
hhsizeXS3 ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m
*Reduce:
#delimit;
```

```
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_s roof_t roof_c wall_b
boat_e cellphone_e h_lit_e plit_e reg3 tonlesap plnmount
hhsizeXS3 ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom
car cellphone electric motorbike tv
floor_s roof_t wall_b
boat_e cellphone_e h_lit_e plit_e reg3 tonlesap plnmount
hhsizeXS3 ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom car cellphone electric motorbike
tv floor_s wall_b boat_e cellphone_e h_lit_e plit_e reg3
```

```
tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom car cellphone electric motorbike
tv wall_b boat_e cellphone_e h_lit_e plit_e reg3
tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit pseced notoilet numroom car cellphone motorbike
tv wall_b boat_e cellphone_e h_lit_e plit_e reg3
tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
```

```
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit notoilet numroom car cellphone motorbike
tv wall_b boat_e cellphone_e h_lit_e plit_e reg3
tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit numroom car cellphone motorbike  tv wall_b boat_e
cellphone_e h_lit_e plit_e reg3 tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"
```

```
*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit numroom car cellphone motorbike  tv boat_e
cellphone_e h_lit_e plit_e reg3 tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit numroom cellphone motorbike  tv boat_e
cellphone_e h_lit_e plit_e reg3 tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit numroom cellphone motorbike  tv
cellphone_e h_lit_e plit_e reg3 tonlesap plnmount hhsizeXS3";
#delimit cr
```

```
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit cellphone motorbike  tv
cellphone_e h_lit_e plit_e reg3 tonlesap plnmount hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit cellphone motorbike  tv
cellphone_e h_lit_e plit_e reg3 tonlesap hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
```

```
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap
*Reduce:
#delimit;
global xvar "hhsize lnhhsz plit cellphone motorbike  tv
cellphone_e h_lit_e plit_e reg3 hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit
*Reduce:
#delimit;
global xvar "hhsize lnhhsz cellphone motorbike  tv cellphone_e h_lit_e plit_e reg3
hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e
*Reduce:
#delimit;
```

```
global xvar "hhsize lnhhsz cellphone motorbike  tv cellphone_e h_lit_e reg3
hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e
*Reduce:
#delimit;
global xvar "hhsize lnhhsz cellphone motorbike tv cellphone_e  reg3 hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"


*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike
*Reduce:
#delimit;
global xvar "hhsize lnhhsz cellphone  tv cellphone_e reg3 hhsizeXS3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
```

```
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike hhsizeXS3
*Reduce:
#delimit;
global xvar "hhsize lnhhsz cellphone  tv cellphone_e reg3";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike hhsizeXS3 reg3
*Reduce:
#delimit;
global xvar "hhsize lnhhsz cellphone  tv cellphone_e ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike hhsizeXS3 reg3
*hhsize
*Reduce:
```

```
#delimit;
global xvar " lnhhsz cellphone  tv cellphone_e ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike hhsizeXS3 reg3
*hhsize tv
*Reduce:
#delimit;
global xvar "lnhhsz cellphone cellphone_e ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike hhsizeXS3 reg3
*hhsize tv lnhhsize
*Reduce:
#delimit;
global xvar "cellphone cellphone_e ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
```

```
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*Removing Phone and Radio roof_cXS3 pkids06 rfree computer motorbikeXS3 floor_c
*floor_t resplus_e numroomXS3 roof_m roof_c roof_t floor_s electric pseced notoilet wall_b
*car boat_e numroom plnmount tonlesap plit plit_e h_lit_e motorbike hhsizeXS3 reg3
*hhsize tv lnhhsz cellphone
*Reduce:
#delimit;
global xvar "cellphone_e ";
#delimit cr
svy: reg ln_exp $xvar
* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*"$panal\RanDraw_red.do"-This is adpted from Haslett et al(2013)
/* Random beta  */
drop _all
global px1=$px+1
set obs $px1
gen z=invnorm(uniform())
mkmat z
matrix bstar=beta+Vbh*z

/* Random alpha  */
drop _all
global pz1=$pz+1
set obs $pz1
gen z=invnorm(uniform())
mkmat z
matrix astar=alpha+Vah*z

/* Loops done differently in Stata10 */

use $cnew\CensusExpS_v1_red, clear
global censC=psuc[_N]
```

```
gen xb=bstar[$px1,1]
forvalues i = 1/$px {
local v : word `i' of $xvar
replace xb=xb+bstar[`i',1]*`v'
}
gen za=astar[$pz1,1]
forvalues i = 1/$pz {
local v : word `i' of $zvar
replace za=za+astar[`i',1]*`v'
}
gen Bb=exp(za)
save "$temp\CENSUSw", replace  /*Working copy of census file*/

/* Sample psus in PSUerr*/
use "$snew\PSUerr", clear
if _N<$censC set obs $censC
gen rc=int(uniform()*$survC)+1
save "$temp\PSUerrb", replace

/* Now merge with HHerr1 and PSUerrb, to construct Y*/
use "$temp\CENSUSw", replace
merge using "$temp\PSUerrb"
gen psub=rc[psuc]
gen hb=hi[psub]
gen ncb=nc[psub]
gen Ncb=Nc[psub]
drop _merge

/* Draw estar from within cluster chosen for hi */
/* (hence peculiar messing about with ncb)      */
merge using "$snew\HHerr1"
gen rn=Ncb-int(uniform()*ncb)
gen estarb=estar[rn]
gen sdb=sqrt(($A*Bb-0.0001)/(1+Bb)+$s2r/2*($A+0.0001)*Bb*(1-Bb)/(1+Bb)^3)
gen eb=estarb*sdb
drop _merge

gen Yb$j=exp(xb+eb+hb)
drop psuc-eb

sort ic
save "$temp\CENSUSw", replace
use "$results\CensusExp_r1", clear
sort ic
```

```
merge ic using "$temp\CENSUSw"
drop _merge
save "$results\CensusExpS_r_red", replace
*each time the model is run the results are saved to "results\CensusExpS_r_red".

\*From here the SAE estimates are generated at each level of aggregation where *r denotes
how many variables are removed from the reduced model.
forvalues iR=1/5   {
use $results\CensusExp_r‘iR’.dta, clear
*replace pline=193052*12/365
*replace pline=132386*12/365 if reg2==1
*replace pline=106560*12/365 if reg3==1
drop if Yb1==.
local i=1
while ‘i’<=100 {
replace Yb‘i’=(Yb‘i’<pline)
local i=‘i’+1
}
compress
collapse (count) size=Yb1 (mean) Yb1-Yb100 [pweight=hhsize], by(psuc) fast
sort psuc
merge psuc, using $cnew\CensusExp_areas‘iR’.dta
drop _merge
}
if ‘iR’>1   {
append using $results\P0ea_p*r.dta
save $results\P0ea_p*r.dta, replace
}
}

compress
sort psuc
save $results\P0ea_p*r.dta, replace

*Commune level:
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], by(province district commune) fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort province district commune
save $results\P0commune*r.dta, replace
*use $results\P0commune*r.dta, clear

*District level:
```

```
use $results\P0ea_p*r.dta, clear
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], by(province district) fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort province district
save $results\P0district*r.dta, replace


*Province level:
use $results\P0ea_p*r.dta, clear
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], by(province) fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort province
save $results\P0province*r.dta, replace


*Region level:
use $results\P0ea_p*r.dta, clear
gen region=1+reg2+2*reg3
label define region 1 "Phnom Penh" 2 "Other urban" 3 "Rural"
label values region region
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], by(region) fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort region
save $results\P0region*r.dta, replace


*Ezone level:
use $results\P0ea_p*r.dta, clear
label define ezone 1 "Phnom Penh" 2 "Plain" 3 "Tonlesap" 4 "Pl/Mntn" 5 "Coastal"
label values ezone ezone
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], by(ezone) fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort ezone
save $results\P0ezone*r.dta, replace
```

```
*Urbanrural level:
use $results\P0ea_p*r.dta, clear
label define rural 0 "Urban" 1 "Rural"
label values rural rural
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], by(rural) fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort rural
save $results\P0rural*r.dta, replace


*Country level:
use $results\P0ea_p*r.dta, clear
collapse (count) Npp=Yb1 (mean) Yb1-Yb100 [pweight=size], fast
egen P0=rmean(Yb1-Yb100)
egen se0=rsd(Yb1-Yb100)
drop Yb1-Yb100
save $results\P0Cambodia*r.dta, replace


use $results\P0district*r.dta, clear
outsheet using $results\P0district*r.csv, comma replace
use $results\P0commune*r.dta, clear
outsheet using $results\P0commune*r.csv, comma replace


*Combining the data files:
forvalues x=0/34 {
use $results\P0commune`x', clear
gen commid=10000*province+100*district+commune
rename P0 p0`x'
rename se0 se`x'
save $results\P0commune`x', replace
}

*use $results\fullmodel, clear
use $results\P0commune0
merge 1:1 commid using $results\P0commune1
drop _merge
merge 1:1 commid using $results\P0commune2
drop _merge
merge 1:1 commid using $results\P0commune3
drop _merge
```

```
merge 1:1 commid using $results\P0commune4
drop _merge
merge 1:1 commid using $results\P0commune5
drop _merge
merge 1:1 commid using $results\P0commune6
drop _merge
merge 1:1 commid using $results\P0commune7
drop _merge
merge 1:1 commid using $results\P0commune8
drop _merge
merge 1:1 commid using $results\P0commune9
drop _merge
merge 1:1 commid using $results\P0commune10
drop _merge
merge 1:1 commid using $results\P0commune11
drop _merge
merge 1:1 commid using $results\P0commune12
drop _merge
merge 1:1 commid using $results\P0commune13
drop _merge
merge 1:1 commid using $results\P0commune14
drop _merge
merge 1:1 commid using $results\P0commune15
drop _merge
merge 1:1 commid using $results\P0commune16
drop _merge
merge 1:1 commid using $results\P0commune17
drop _merge
merge 1:1 commid using $results\P0commune18
drop _merge
merge 1:1 commid using $results\P0commune19
drop _merge
merge 1:1 commid using $results\P0commune20
drop _merge
merge 1:1 commid using $results\P0commune21
drop _merge
merge 1:1 commid using $results\P0commune22
drop _merge
merge 1:1 commid using $results\P0commune23
drop _merge
merge 1:1 commid using $results\P0commune24
drop _merge
merge 1:1 commid using $results\P0commune25
drop _merge
```

```
merge 1:1 commid using $results\P0commune26
drop _m
merge 1:1 commid using $results\P0commune27
drop _merge
merge 1:1 commid using $results\P0commune28
drop _merge
merge 1:1 commid using $results\P0commune29
drop _merge
merge 1:1 commid using $results\P0commune30
drop _merge
merge 1:1 commid using $results\P0commune31
drop _merge
merge 1:1 commid using $results\P0commune32
drop _merge
merge 1:1 commid using $results\P0commune33
drop _merge
merge 1:1 commid using $results\P0commune34
drop _merge
save $results\fullmodel, replace

order

use $results\fullmodel, clear

forvalues i=1/34{
rename se0‘i’ se‘i’
}

order province district commune commid Npp p0 se0 p01 se1 p02 se2 p03 se3 /*
*/p04 se4 p05 se5 p06 se6 p07 se7 p08 se8 p09 se9 p010 se10 p011 se11 p012 se12 /*
*/p013 se13 p014 se14 p015 se15 p016 se16 p017 se17 p018 se18 p019 se19 p020 se20 /*
*/p021 se21 p022 se22 p023 se23 p024 se24 p025 se25 p026 se26 p027 se27 p028 se28 /*
*/p029 se29 p030 se30 p031 se31 p032 se32 p033 se33 p034 se34

forvalues i=1/34 {
gen z‘i’=(p0-p0‘i’)/sqrt((se0^2)+(se‘i’^2))
}


save $results\fullmodel, replace

save "E:/PhD/Cambodia_2017/cambodia_data.csv", replace

* The model plots
```

```
use $results\fullmodel, clear
*Calculating the Spearman and the Pearson correlation
spearman p*
pwcorr p*


order province-rural

if ‘iR’==1   {
save $results\P0ea_p*r.dta, replace

*This is the code from R to generate plots
library(readr)
cambodia <- read_csv("E:/PhD/Cambodia_2017/cambodia_data.csv")
attach(cambodia)
mynames<-c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15",
"16","17","18","19","20","21","22","23","24","25","26",
"27","28","29","30","31","32","33","34")

mynames1<-c("0", "1","2","3","4","5","6","7","8","9","10","11","12","13","14","15",
"16","17","18","19","20","21","22","23","24","25","26",
"27","28","29","30","31","32","33","34")

#z-statistic plot
dev.new(width=8, height=4)
boxplot(z1,z2,z3,z4,z5,z6,z7,z8,z9,z10,z11,z12,z13,z14,z15,
z16,z17,z18,z19,z20,z21,z22,z23,z24,z25,z26,z27,z28
,z29,z30,z31,z32,z33,z34, type=’o’, pch=19, cex=0.5,
cex.axis=0.6 ,names=mynames, xlab="Number of Variables Removed",
ylab="Z-statistic")

#poverty plot
dev.new(width=8, height=4)
boxplot(p0,p01,p02,p03,p04,p05,p06,p07,p08,p09,p010,p011,p012,p013,p014,p015,
p016,p017,p018,p019,p020,p021,p022,p023,p024,p025,p026,p027,p028
,p029,p030,p031,p032,p033,p034, type=’o’, pch=19, cex=0.5,
cex.axis=0.6 ,names=mynames1, xlab="Number of Variables Removed",
ylab="Poverty Incidence")

#standard error plot
dev.new(width=8, height=4)
boxplot(se0, se1,se2,se3,se4,se5,se6,se7,se8,se9,se10,se11,se12,se13,
se14,se15, se16,se17,se18,se19,se20,se21,se22,se23,se24,se25,
```

```
se26,se27,se28,se29,se30,se31,se32,se33,se34, type='o', pch=19, cex=0.5,
cex.axis=0.6 ,names=mynames1, xlab="Number of Variables Removed",
ylab="Standard Error of Poverty Incidence")

detach(cambodia)


*Chapter 6

*Using the initial cleaned data sets from Haslett et al (2014b) where it was initially discover
Not that some of the files are using different global file paths, as the work was done
*between two different computers.
*Note this is not all the code. The preliminary code is not included.

version 13.1
clear

global origs "E:\PhD\Nepal2013_old\Data\NLSS\Raw"
global origc "E:\PhD\Nepal2013_old\Data\Census\Raw"
global origd "E:\PhD\Nepal2013_old\Data\DHS\Raw"

global outs  "E:\PhD\Nepal2013_old\Data\NLSS\Created"
global outc  "E:\PhD\Nepal2013_old\Data\Census\Created"
global outd  "E:\PhD\Nepal2013_old\Data\DHS\Created"

global outedit "E:\PhD\Nepal2013_old\Wasting_edit"
global analysis  "E:\PhD\Nepal2013_old\Analysis\Malnutrition\Wasting"
global results "E:\PhD\Nepal2013_old\results\Wasting"

global temp  "E:\PhD\Nepal2013_old\Data\Temp"

\*This generates the total and the mean excluding observation ith small area in each
district for one variable, then minusing the observation by the mean */

use "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_mean.dta", clear

*variables in the model
#delimit;
global xvars "ageyr23 girl  terai wat_cwell hage2 flr_con wall_wod wall_bambo
wall_brk Wroof_iron Wroof_tile Wroof_straw Wmax_educ_none Whead_female
Wmax_educ_fem_5to7 Wtoilet_flushseptik Wroof_mud Wtoilet_none Wwater_piped
Wowns_fridge meanht popdens"
#delimit cr
```

```
order ilakaid
gen domain=belt*100+region

*deletion diagnsotics, for the population as a whole-unweighted
use "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_mean.dta", clear
#delimit;
global xvars "ageyr23 girl  terai wat_cwell hage2 flr_con wall_wod wall_bambo
wall_brk Wroof_iron Wroof_tile Wroof_straw Wmax_educ_none Whead_female
Wmax_educ_fem_5to7 Wtoilet_flushseptik Wroof_mud Wtoilet_none Wwater_piped
Wowns_fridge meanht popdens"
#delimit cr

order ilakaid
gen domain=belt*100+region
foreach var of varlist ageyr23-popdens{
egen total_‘var’ = total(‘var’)
egen ‘var’_n = count(‘var’)
gen ‘var’_totalMINUSi = total_‘var’ - cond(missing(‘var’), 0, ‘var’)
gen ‘var’_meanMINUSi = ‘var’_totalMINUSi / (‘var’_n - !missing(‘var’))
gen ‘var’_infl=‘var’-‘var’_meanMINUSi
drop total_‘var’ ‘var’_totalMINUSi
}

foreach var of varlist ageyr23-popdens{
drop ‘var’_infl ‘var’_n
}

save "E:\PhD\Nepal2013_old\Wasting_edit\CensusWa_influence.dta", replace

use "$outd\dhs_gis", clear

drop if ZHW>5 | ZHW<-5

replace hhwt=hhwt/1000000

gen hhszsq=(hhsize-6)^2

//collapsing of categories
gen ageyr23=ageyr2+ageyr3
gen ageyr45=ageyr4+ageyr5

gen wall_brk=wall_cmtbrk+wall_mudbrk+wall_ubrk

rename ZHW whz
```

```
svyset psu [pweight=hhwt], strata(strat_des)
svy: regress whz $xvar
matrix beta=e(b)'
mat betaT=beta'

use "E:\PhD\Nepal2013_old\Wasting_edit\CensusWa_influence.dta", clear

svmat betaT
rename betaT1 beta_ageyr23
rename betaT2 beta_girl
rename betaT3 beta_terai
rename betaT4 beta_wat_cwell
rename betaT5 beta_hage2
rename betaT6 beta_flr_con
rename betaT7 beta_wall_wod
rename betaT8 beta_wall_bambo
rename betaT9 beta_wall_brk
rename betaT10 beta_Wroof_iron
rename betaT11 beta_Wroof_tile
rename betaT12 beta_Wmax_educ_none
rename betaT13 beta_Whead_female
rename betaT14 beta_Wroof_straw
rename betaT15 beta_Wmax_educ_fem_5to7
rename betaT16 beta_Wtoilet_flushseptik
rename betaT17 beta_Wroof_mud
rename betaT18 beta_Wtoilet_none
rename betaT19 beta_Wwater_piped
rename betaT20 beta_Wowns_fridge
rename betaT21 beta_meanht
rename betaT22 beta_popdens

foreach var of varlist beta_ageyr23-beta_popdens{
replace `var' = `var'[_n-1] if missing(`var')
}

foreach var of varlist ageyr23-popdens{
gen `var'_bd=`var'_infl*beta_`var'
}

save "E:\PhD\Nepal2013_old\Wasting_edit\CensusWa_influence.dta", replace

use "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_mean.dta", clear
order ilakaid
```

```
gen domain=belt*100+region

*Loop for all the variables, with no weights for the localised deletion diagnostic
foreach var of varlist ageyr23-popdens{
egen total_‘var’ = total(‘var’), by(dcode)
egen ‘var’_n = count(‘var’), by(dcode)
gen ‘var’_totalMINUSi = total_‘var’ - cond(missing(‘var’), 0, ‘var’)
gen ‘var’_meanMINUSi = ‘var’_totalMINUSi / (‘var’_n - !missing(‘var’))
gen ‘var’_d=‘var’-‘var’_meanMINUSi
drop total_‘var’ ‘var’_n ‘var’_totalMINUSi ‘var’_meanMINUSi
}

foreach var of varlist ageyr23-popdens{
drop total_‘var’ ‘var’_n ‘var’_totalMINUSi ‘var’_meanMINUSi
}

save "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_inf.dta", replace
use "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_inf.dta", clear

matrix beta=e(b)’
mat list beta
mat betaT=beta’

svmat beta
rename betaT1 beta_ageyr23
rename betaT2 beta_girl
rename betaT3 beta_terai
rename betaT4 beta_wat_cwell
rename betaT5 beta_hage2
rename betaT6 beta_flr_con
rename betaT7 beta_wall_wod
rename betaT8 beta_wall_bambo
rename betaT9 beta_wall_brk
rename betaT10 beta_Wroof_iron
rename betaT11 beta_Wroof_tile
rename betaT12 beta_Wmax_educ_none
rename betaT13 beta_Whead_female
rename betaT14 beta_Wroof_straw
rename betaT15 beta_Wmax_educ_fem_5to7
rename betaT16 beta_Wtoilet_flushseptik
rename betaT17 beta_Wroof_mud
rename betaT18 beta_Wtoilet_none
rename betaT19 beta_Wwater_piped
rename betaT20 beta_Wowns_fridge
```

```
rename betaT21 beta_meanht
rename betaT22 beta_popdens

rename beta1 beta_ageyr23
rename beta2 beta_girl
rename beta3 beta_terai
rename beta4 beta_wat_cwell
rename beta5 beta_hage2
rename beta6 beta_flr_con
rename beta7 beta_wall_wod
rename beta8 beta_wall_bambo
rename beta9 beta_wall_brk
rename beta10 beta_Wroof_iron
rename beta11 beta_Wroof_tile
rename beta12 beta_Wmax_educ_none
rename beta13 beta_Whead_female
rename beta14 beta_Wroof_straw
rename beta15 beta_Wmax_educ_fem_5to7
rename beta16 beta_Wtoilet_flushseptik
rename beta17 beta_Wroof_mud
rename beta18 beta_Wtoilet_none
rename beta19 beta_Wwater_piped
rename beta20 beta_Wowns_fridge
rename beta21 beta_meanht
rename beta22 beta_popdens

foreach var of varlist beta_ageyr23-beta_popdens{
replace `var' = `var'[_n-1] if missing(`var')
}

foreach var of varlist ageyr23-popdens{
gen `var'_bd=`var'_d*beta_`var'
}

save "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_inf_dist.dta", replace

*weighted deletion diagnsostics

use "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_mean.dta", clear

*global level deletion diagnostics
gen domain=belt*100+region
rename popsize Nhh
egen TNhh=total(Nhh)
```

```
foreach var of varlist ageyr23-popdens{
gen T‘var’=‘var’*Nhh
egen total_‘var’ = total(T‘var’)
gen ‘var’_totalMINUSi = total_‘var’ -  T‘var’
gen ‘var’_meanMINUSi = ‘var’_totalMINUSi / (TNhh - Nhh)
gen ‘var’_infl=‘var’-‘var’_meanMINUSi
drop T‘var’ total_‘var’ ‘var’_totalMINUSi ‘var’_meanMINUSi
}

svmat betaT
rename betaT1 beta_ageyr23
rename betaT2 beta_girl
rename betaT3 beta_terai
rename betaT4 beta_wat_cwell
rename betaT5 beta_hage2
rename betaT6 beta_flr_con
rename betaT7 beta_wall_wod
rename betaT8 beta_wall_bambo
rename betaT9 beta_wall_brk
rename betaT10 beta_Wroof_iron
rename betaT11 beta_Wroof_tile
rename betaT12 beta_Wroof_straw
rename betaT13 beta_Wmax_educ_none
rename betaT14 beta_Whead_female
rename betaT15 beta_Wmax_educ_fem_5to7
rename betaT16 beta_Wtoilet_flushseptik
rename betaT17 beta_Wroof_mud
rename betaT18 beta_Wtoilet_none
rename betaT19 beta_Wwater_piped
rename betaT20 beta_Wowns_fridge
rename betaT21 beta_meanht
rename betaT22 beta_popdens

foreach var of varlist beta_ageyr23-beta_popdens{
replace ‘var’ = ‘var’[_n-1] if missing(‘var’)
}

foreach var of varlist ageyr23-popdens{
gen ‘var’_bd=‘var’_infl*beta_‘var’
}

save "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_infl_weighted.dta", replace

drop ageyr23-popdens
```

```
rename (ageyr23_bd girl_bd terai_bd wat_cwell_bd hage2_bd /*
*/flr_con_bd wall_wod_bd wall_bambo_bd wall_brk_bd Wroof_iron_bd/*
*/ Wroof_tile_bd Wroof_straw_bd Wmax_educ_none_bd Whead_female_bd /*
*/Wmax_educ_fem_5to7_bd Wtoilet_flushseptik_bd Wroof_mud_bd Wtoilet_none_bd/*
*/ Wwater_piped_bd Wowns_fridge_bd meanht_bd popdens_bd) /*
*/(ageyr23 girl terai wat_cwell hage2 /*
*/flr_con wall_wod wall_bambo wall_brk Wroof_iron /*
*/Wroof_tile Wroof_straw Wmax_educ_none Whead_female /*
*/Wmax_educ_fem_5to7 Wtoilet_flushseptik Wroof_mud Wtoilet_none/*
*/ Wwater_piped Wowns_fridge meanht popdens)

*this is creating Fig 6.5
*boxplot for global influence statistic for the weighted data
graph box ageyr23- popdens, showyvars yvaroptions(label(angle(vertical)/*
*/labsize(vsmall))) marker(1, msize(vsmall)) marker(2,/*
*/ msize(vsmall)) marker(3, msize(vsmall)) marker(4, msize(vsmall)) marker(5,/*
*/ msize(vsmall)) marker(6, msize(vsmall)) marker(7, msize(vsmall)) marker(/*
*/ 8, msize(vsmall)) marker(9, msize(vsmall)) marker(10, msize(vsmall)) marker(11,/*
*/ msize(vsmall)) marker(12, msize(vsmall)) marker(13, msize(vsmall)) marker(14,/*
*/ msize(vsmall)) marker(15, msize(vsmall)) marker(16, msize(vsmall)) /*
*/marker(17, msize(vsmall)) marker(18, msize(vsmall)) marker(19, msize(vsmall))/*
*/marker(20, msize(vsmall)) marker(21, msize(vsmall)) marker(22, msize(vsmall))/*
*/ ytitle(influence diagnostics) subtitle(, size(vsmall) span) /*
*/legend(off)

*localised weighted deletion diagnostic
use "E:\PhD\Nepal2013_old\Wasting_edit\CensusWa_mean.dta", clear

use "$outd\dhs_gis", clear

drop if ZHW>5 | ZHW<-5

replace hhwt=hhwt/1000000

gen hhszsq=(hhsize-6)^2

//collapsing of categories
gen ageyr23=ageyr2+ageyr3
gen ageyr45=ageyr4+ageyr5

gen wall_brk=wall_cmtbrk+wall_mudbrk+wall_ubrk
```

```
#delimit;
global xvar "ageyr23 girl terai wat_cwell hage2 flr_con
wall_wod wall_bambo wall_brk Wroof_iron Wroof_tile Wroof_straw  Wmax_educ_none
Whead_female  Wmax_educ_fem_5to7 Wtoilet_flushseptik  Wroof_mud
Wtoilet_none  Wwater_piped Wowns_fridge meanht popdens ";
#delimit cr

rename ZHW whz

svyset psu [pweight=hhwt], strata(strat_des)
svy: regress whz $xvar
matrix beta=e(b)'
mat betaT=beta'


gen domain=belt*100+region
rename popsize Nhh
egen TNhh=total(Nhh)
foreach var of varlist ageyr23-popdens{
gen T`var'=`var'*Nhh, by(dcode)
egen dtotal_`var' = total(T`var')
gen `var'_totalMINUSi = total_`var' -  T`var'
gen `var'_meanMINUSi = `var'_totalMINUSi / (TNhh - Nhh)
gen `var'_infl=`var'-`var'_meanMINUSi
drop T`var' total_`var' `var'_totalMINUSi `var'_meanMINUSi
}

svmat betaT
rename betaT1 beta_ageyr23
rename betaT2 beta_girl
rename betaT3 beta_terai
rename betaT4 beta_wat_cwell
rename betaT5 beta_hage2
rename betaT6 beta_flr_con
rename betaT7 beta_wall_wod
rename betaT8 beta_wall_bambo
rename betaT9 beta_wall_brk
rename betaT10 beta_Wroof_iron
rename betaT11 beta_Wroof_tile
rename betaT12 beta_Wroof_straw
rename betaT13 beta_Wmax_educ_none
rename betaT14 beta_Whead_female
rename betaT15 beta_Wmax_educ_fem_5to7
rename betaT16 beta_Wtoilet_flushseptik
```

```
rename betaT17 beta_Wroof_mud
rename betaT18 beta_Wtoilet_none
rename betaT19 beta_Wwater_piped
rename betaT20 beta_Wowns_fridge
rename betaT21 beta_meanht
rename betaT22 beta_popdens

foreach var of varlist beta_ageyr23-beta_popdens{
replace ‘var’ = ‘var’[_n-1] if missing(‘var’)
}

foreach var of varlist ageyr23-popdens{
gen ‘var’_bd=‘var’_infl*beta_‘var’
}

save "E:\PhD\Nepal2013_old\Data\Census\Created\CensusWa_infl_weighted_district.dta", replace
drop ageyr23-popdens

rename (ageyr23_bd girl_bd terai_bd wat_cwell_bd hage2_bd /*
*/flr_con_bd wall_wod_bd wall_bambo_bd wall_brk_bd Wroof_iron_bd/*
*/ Wroof_tile_bd Wroof_straw_bd Wmax_educ_none_bd Whead_female_bd /*
*/Wmax_educ_fem_5to7_bd Wtoilet_flushseptik_bd Wroof_mud_bd Wtoilet_none_bd/*
*/ Wwater_piped_bd Wowns_fridge_bd meanht_bd popdens_bd) /*
*/(ageyr23 girl terai wat_cwell hage2 /*
*/flr_con wall_wod wall_bambo wall_brk Wroof_iron /*
*/Wroof_tile Wroof_straw Wmax_educ_none Whead_female /*
*/Wmax_educ_fem_5to7 Wtoilet_flushseptik Wroof_mud Wtoilet_none/*
*/ Wwater_piped Wowns_fridge meanht popdens)

*Fig 6.6
*boxplot for global influence statistic for the weighted data
graph box ageyr23- popdens, showyvars yvaroptions(label(angle(vertical)/*
*/labsize(vsmall))) marker(1, msize(vsmall)) marker(2,/*
*/ msize(vsmall)) marker(3, msize(vsmall)) marker(4, msize(vsmall)) marker(5,/*
*/ msize(vsmall)) marker(6, msize(vsmall)) marker(7, msize(vsmall)) marker(/*
*/ 8, msize(vsmall)) marker(9, msize(vsmall)) marker(10, msize(vsmall)) marker(11,/*
*/ msize(vsmall)) marker(12, msize(vsmall)) marker(13, msize(vsmall)) marker(14,/*
*/ msize(vsmall)) marker(15, msize(vsmall)) marker(16, msize(vsmall)) /*
*/marker(17, msize(vsmall)) marker(18, msize(vsmall)) marker(19, msize(vsmall))/*
*/marker(20, msize(vsmall)) marker(21, msize(vsmall)) marker(22, msize(vsmall))/*
*/ ytitle(influence diagnostics(Distrtict level)) subtitle(, size(vsmall) span) /*
*/legend(off)
```

```
\*This is producing Figure 6.3 (This is simialr to how to produce 6.2-so haven't included the c
*stacking the data
*use "$outedit\census_district_diff", clear
gen domain=dcode

stack ilakaid domain ageyr23_bd  ilakaid domain girl_bd  ilakaid domain terai_bd/*
*/  ilakaid domain wat_cwell_bd  ilakaid domain wat_cwell_bd  ilakaid domain flr_con_bd/*
*/  ilakaid domain wall_wod_bd  ilakaid domain wall_bambo_bd  ilakaid domain wall_brk_bd/*
*/  ilakaid domain Wroof_iron_bd  ilakaid domain Wroof_tile_bd  ilakaid domain /*
*/Wroof_straw_bd  ilakaid domain Wmax_educ_none_bd  ilakaid domain Whead_female_bd/*
*/  ilakaid domain Wmax_educ_fem_5to7_bd  ilakaid domain Wtoilet_flushseptik_bd/*
*/  ilakaid domain Wroof_mud_bd  ilakaid domain Wtoilet_none_bd  ilakaid domain/*
*/ Wwater_piped_bd  ilakaid domain Wowns_fridge_bd  ilakaid domain meanht_bd/*
*/  ilakaid domain popdens_bd,  into(ilakid domain influence) clear wide


*label drop variable_lab
gen variable=_stack
label define variable_lab 1 ageyr23 2 girl 3 terai 4 wat_cwell 5 wat_cwell /*
*/6 flr_con 7 wall_wod 8 wall_bambo 9 wall_brk 10 Wroof_iron /*
*/11 Wroof_tile 12 Wroof_straw 13 Wmax_educ_none 14 Whead_female /*
*/15 Wmax_educ_fem_5to7 16 Wtoilet_flushseptik 17 Wroof_mud 18 Wtoilet_none/*
*/ 19 Wwater_piped 20 Wowns_fridge 21 meanht 22 popdens
label values variable variable_lab

gen ilaka901=.
replace ilaka901 = 1 if ilakid==901
replace ilaka901 = 0 if ilaka901==.
*droping popdens due to the hug variation
*drop if _stack==22

twoway (scatter influence variable if ilaka901==0, mcolor(blue) msize(vsmall)) (scatter /*
*/influence variable if ilaka901==1, mcolor(red) msize(vsmall)), xlabel(#21, labels /*
*/labsize(small) angle(vertical) format(%12s) valuelabel) legend(off)

graph export "E:\PhD\Nepal2013_old\Wasting_edit\influence_district_mean_small.pdf", /*
*/as(pdf) replace
save "$outedit\stacked_districtmean_ilaka", replace

save "E:\PhD\Nepal2013_old\Wasting_edit\CensusWa_influence_beta.dta", replace

*Found that Wroof type is unusual so checking this*

checking the mean of the Wroof_straw and Wroof_other
```

```
use $origc\CensusMeans_ward, clear
rename district dcode
rename vdc vcode
merge m:1 dcode vcode ward using "$origd\ilakaid"
keep if _merge==3
drop _merge
gen ilakaid=100*dcode+ilaka
keep if ilakaid==901

*This is looking at ilaka 901 and how roof type is behaving
*this is taking the mean at person level, should be weighted by household
summarize Wroof_straw Wroof_other [fweight = popn]
by vcode, sort : summarize Wroof_straw Wroof_other [fweight = popn]
save $outc\CensusMeans_ilaka901, replace

use F:\Nepal2013\Maris_files\Data\Census\Created\HHAmen_c, clear
gen vdcode=vdcmun
collapse (mean) Wroof_straw=roof_straw Wroof_galv=roof_galv/*
*/ Wroof_tile=roof_tile Wroof_conc=roof_conc Wroof_wod=roof_wod Wroof_oth=/*
*/roof_oth (count) numhh=vdcode, by(dist vdcmun ward)


use F:\Nepal2013_old\Maris_files\Data\Census\Created\HHAmen_c, clear
collapse (count) numhh=hno, by(dist vdcmun ward)
rename dist dcode
rename vdcmun vcode
merge m:1 dcode vcode ward using "$origd\ilakaid"
keep if _merge==3
gen ilakaid=100*dcode+ilaka
keep if ilakaid==901
drop _merge
save $outc\hholds_ilaka901, replace
use $outc\hholds_ilaka901, clear

use $outc\CensusMeans_ilaka901_roof, clear
merge 1:1 dcode vcode ward using $outc\hholds_ilaka901
save $outc\CensusMeans_ilaka901_roof, replace
drop _merge
drop ilaka urbrurl belt region batchid9
order ilakaid dcode vcode ward batchid popn numhh


*Having found that there is an error variable Wroof_straw in ilaka 901, need to rerun the analy
```

```
*Master file
version 13.1
clear

global origs "F:\Nepal2013_old\Data\NLSS\Raw"
global origc "F:\Nepal2013_old\Data\Census\Raw"
global origd "F:\Nepal2013_old\Data\DHS\Raw"

global outs  "F:\Nepal2013_old\Data\NLSS\Created"
global outc  "F:\Nepal2013_old\Data\Census\Created"
global outd  "F:\Nepal2013_old\Data\DHS\Created"

global analysis  "F:\Nepal2013_old\Analysis\Malnutrition\Wasting"
global results "F:\Nepal2013_old\results\Wasting"

global outedit "F:\Nepal2013_old\Wasting_edit"

global temp  "F:\Nepal2013_old\Data\Temp"

use $outedit\ilakaid_r901, clear
drop Yb*
save $outedit\ilakaid_r901, replace

run $analysis\SurvReg_f.do

global j=1
while $j<=100                          {
run "$outedit\RanDraw_edit.do"
global j=$j+1
}

*This is the $analysis\SurvReg_f.do-This file was created in Haslett et al (2014b)
use "$outd\dhs_gis", clear

drop if ZHW>5 | ZHW<-5

replace hhwt=hhwt/1000000

gen hhszsq=(hhsize-6)^2

//collapsing of categories
gen ageyr23=ageyr2+ageyr3
gen ageyr45=ageyr4+ageyr5
```

```
gen wall_brk=wall_cmtbrk+wall_mudbrk+wall_ubrk


#delimit;
global xvar "ageyr23 girl terai wat_cwell hage2 flr_con
wall_wod wall_bambo wall_brk Wroof_iron Wroof_tile  Wmax_educ_none
Whead_female Wroof_straw Wmax_educ_fem_5to7 Wtoilet_flushseptik  Wroof_mud
Wtoilet_none  Wwater_piped Wowns_fridge meanht popdens ";
#delimit cr

rename ZHW whz

svyset psu [pweight=hhwt], strata(strat_des)
svy: regress whz $xvar


* Things to keep
predict r, resid
egen hhs=group(hhid)
egen psuid=group(psu)
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
xtmixed r ||psuid: ||hhs:
matrix sigmat=e(b)'
scalar sigc=exp(sigmat[2,1])
scalar sigh=exp(sigmat[3,1])
scalar sige=exp(sigmat[4,1])


*The $outedit\RanDraw_edit.do is:
/* Prediction of Yb for census data                                    */
/* Predictions Yb1-Yb100 into $Rname.dta                               */
/* starting with census data in $CensusSt_vx.dta                      */

/* This version uses a single model, kept in SurvReg_f     */


/* Random beta  */
drop _all
global px1=$px+1
set obs $px1
gen z=invnorm(uniform())
```

```
mkmat z
matrix bstar=beta+Vbh*z

/* Each region file done separately to minimize memory restrictions */
use "$outedit\ilaka901", clear
*drop Troof_straw Wroof_straw
by psuc, sort: gen psucnew = _n == 1
replace psucnew = sum(psucnew)
by hhc, sort: gen hhcnew = _n == 1
replace hhcnew = sum(hhcnew)
drop Wroof_straw
gen xran=rbinomial(Troof_other,rbeta(98,9))
replace xran=0 if xran==.
by psuc, sort: gen var1=_n==1
gen xran2=var1*xran
replace xran2 = xran2[_n-1] if xran2==0

gen Troof_straw_ran=xran2+Troof_straw
gen Wroof_straw=Troof_straw_ran/numhh
gen xb=bstar[$px1,1]

forvalues i = 1/$px {
local v : word `i' of $xvar
replace xb=xb+bstar[`i',1]*`v'
}
gen eb=invnorm(uniform())*sige
gen zh=invnorm(uniform())*sigh
gen zc=invnorm(uniform())*sigc

gen cb=zc[psucnew]
gen hb=zh[hhcnew]

gen Yb$j=xb+cb+hb+eb

keep id Yb$j
sort id
save "$temp\CENSUSw901", replace
use "$outedit\ilakaid_r901", clear
sort id
merge id using "$temp\CENSUSw901"
drop _merge
save "$outedit\ilakaid_r901", replace

use "$outedit\ilakaid_r901", clear
```

```
sort batchid
drop if batchid==.
save "$outedit\ilakaid_r901", replace

/* This is generating the new Wasting estimates (W2) for ilaka 901*/

global waline=-2

use "$outedit\ilakaid_r901", clear
keep if Yb1<.
local i=1
while `i'<=100 {
replace Yb`i'=(Yb`i'<$waline)
local i=`i'+1
}
compress
sort id
collapse (count) size=Yb1 (mean) belt region urbrurl Yb1-Yb100, by(batchid) fast
save "$outedit\W2ward_p901", replace

sort batchid
merge batchid using "$origd\Ilakaid", keep(urbrurl ilaka dcode vcode)
keep if _merge==3
drop _m
save "$outedit\W2ward_p901", replace

*vdc level:
collapse (count) Npp=Yb1 (mean) region Yb1-Yb100 [pw=size], by(dcode vcode) fast
egen W2=rmean(Yb1-Yb100)
egen se2=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort dcode vcode
save "$outedit\W2vdc901", replace

*Ilaka level:
use "$outedit\W2ward_p901", clear
collapse (count) Npp=Yb1 (mean) region Yb1-Yb100 [pw=size], by(dcode ilaka) fast
egen W2=rmean(Yb1-Yb100)
egen se2=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort dcode ilaka
save "$outedit\W2ilaka901", replace

/*This is generating the new severe wasting estimates (W3) for ilaka 901*/
```

```
global waline=-3

use "$outedit\ilakaid_r901", clear
keep if Yb1<.
local i=1
while `i'<=100 {
replace Yb`i'=(Yb`i'<$waline)
local i=`i'+1
}
compress
sort id
collapse (count) size=Yb1 (mean) belt region urbrurl Yb1-Yb100, by(batchid) fast
save "$outedit\W3ward_p901", replace

sort batchid
merge batchid using "$origd\Ilakaid", keep(urbrurl ilaka dcode vcode)
keep if _merge==3
drop _m
save "$outedit\W3ward_p901", replace


*vdc level:
collapse (count) Npp=Yb1 (mean) region Yb1-Yb100 [pw=size], by(dcode vcode) fast
egen W2=rmean(Yb1-Yb100)
egen se2=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort dcode vcode
save "$outedit\W3vdc901", replace

*Ilaka level:
use "$outedit\W3ward_p901", clear
collapse (count) Npp=Yb1 (mean) region Yb1-Yb100 [pw=size], by(dcode ilaka) fast
egen W2=rmean(Yb1-Yb100)
egen se2=rsd(Yb1-Yb100)
drop Yb1-Yb100
sort dcode ilaka
save "$outedit\W3ilaka901", replace



*Chapter 7

**Generating Figure 7.1
*generating the k to see the Pearson Correlation
```

```
set obs 100
egen n = seq(), f(10) t(100) b(1)
gen k=n*0.1
gen rho= sqrt(k^2-1)/k
twoway (scatter rho k), ytitle(Correlation) xtitle(k) title(Pearson Correlation)
save "E:\PhD\PhDpart3\Simulate\rho_k"

*Simulated data used to generate Figure 7.2 and figure 7.3
*simulation corrected

/*simulation finding the spearman rank correlation between y and yhat at different
ratios of variation between to standard error*/

/*
k=10 e=0.1005
k=9 e=0.1118
k=8 e=0.126
k=7 e=0.1443
k=6 e=0.169
k=5 e=0.204
k=4 e=0.2582
k=3 e=0.3536
k=2 e=0.5773
k=1.01 e=7.053
*/

*spearman correlation
clear all
set seed 12345
set obs 1000
*number of observations (_N) was 0, now 1000
generate y = rnormal(0,1)

*rho10
forvalues j=1/100{
gen se_‘j’=rnormal(0.1005,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho10)
```

```
mat drop rho

*rho9
forvalues j=1/100{
gen se_‘j’=rnormal(0.1118,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho9)
mat drop rho

*rho8
forvalues j=1/100{
gen se_‘j’=rnormal(0.126,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho8)
mat drop rho

*rho7
forvalues j=1/100{
gen se_‘j’=rnormal(0.1443,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho7)
mat drop rho

*rho6
forvalues j=1/100{
gen se_‘j’=rnormal(0.169,1)
```

```
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho6)
mat drop rho

*rho5
forvalues j=1/100{
gen se_‘j’=rnormal(0.204,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho5)
mat drop rho

*rho4
forvalues j=1/100{
gen se_‘j’=rnormal(0.2582,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho4)
mat drop rho

*rho3
forvalues j=1/100{
gen se_‘j’=rnormal(0.3536,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
```

```
}
drop yhat* se*
svmat rho, name(rho3)
mat drop rho

*rho2
forvalues j=1/100{
gen se_‘j’=rnormal(0.5773,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho2)
mat drop rho

*rho1
forvalues j=1/100{
gen se_‘j’=rnormal(7.053,1)
gen yhat‘j’=y+se_‘j’
quietly spearman y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho1)
mat drop rho

rename (rho101 rho91 rho81 rho71 rho61 rho51 rho41 rho31 rho21 rho11) /*
*/(rho10 rho9 rho8 rho7 rho6 rho5 rho4 rho3 rho2 rho1)

save "E:\PhD\PhDpart3\simulation_spear_edit.dta", replace


*correlation

clear all
set seed 12345
set obs 1000
*number of observations (_N) was 0, now 1000
generate y = rnormal(0,1)
```

```
*rho10
forvalues j=1/100{
gen se_‘j’=rnormal(0.1005,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho10)
mat drop rho

*rho9
forvalues j=1/100{
gen se_‘j’=rnormal(0.1118,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho9)
mat drop rho

*rho8
forvalues j=1/100{
gen se_‘j’=rnormal(0.126,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho8)
mat drop rho

*rho7
forvalues j=1/100{
gen se_‘j’=rnormal(0.1443,0)
gen yhat‘j’=y+se_‘j’
```

```
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho7)
mat drop rho

*rho6
forvalues j=1/100{
gen se_‘j’=rnormal(0.169,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho6)
mat drop rho

*rho5
forvalues j=1/100{
gen se_‘j’=rnormal(0.204,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho5)
mat drop rho

*rho4
forvalues j=1/100{
gen se_‘j’=rnormal(0.2582,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
```

```
drop yhat* se*
svmat rho, name(rho4)
mat drop rho

*rho3
forvalues j=1/100{
gen se_‘j’=rnormal(0.3536,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho3)
mat drop rho

*rho2
forvalues j=1/100{
gen se_‘j’=rnormal(0.5773,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho2)
mat drop rho

*rho1
forvalues j=1/100{
gen se_‘j’=rnormal(7.053,0)
gen yhat‘j’=y+se_‘j’
quietly corr y yhat‘j’
scalar r=r(rho)
matrix R=(r)
matrix rho=nullmat(rho)\R
}
drop yhat* se*
svmat rho, name(rho1)
mat drop rho

rename (rho101 rho91 rho81 rho71 rho61 rho51 rho41 rho31 rho21 rho11) /*
```

```
*/(rho10 rho9 rho8 rho7 rho6 rho5 rho4 rho3 rho2 rho1)
save "E:\PhD\PhDpart3\Simulate\simulation_cor_edit.dta", replace

use "E:\PhD\PhDpart3\simulation_spear_edit.dta", clear
order y rho1 rho2 rho3 rho4 rho5 rho6 rho7 rho8 rho9 rho10

*Spearman Correlation
graph box rho1-rho10, showyvars yvaroptions(relabel(1 "1.01" 2 "2" 3 "3" /*
*/4 "4" 5 "5" 6 "6" 7 "7" 8 "8" 9 "9" 10 "10" )/*
*/label(angle(zero) labsize(small))) ytitle(Spearman Correlation)/*
*/title(Spearman Correlation of Simulation) caption(Ratio of Between/*
*/ Small Area Variation to Within Error Variation) legend(off)
graph export "E:\PhD\PhDpart3\Simulate\spearman_edit.pdf", as(pdf) replace

*Pearson Correlation
use "E:\PhD\PhDpart3\simulation_cor_edit.dta", clear
order y rho1 rho2 rho3 rho4 rho5 rho6 rho7 rho8 rho9 rho10

graph box rho1-rho10, showyvars yvaroptions(relabel(1 "1.01" 2 "2" 3 "3" /*
*/4 "4" 5 "5" 6 "6" 7 "7" 8 "8" 9 "9" 10 "10" )/*
*/label(angle(zero) labsize(small))) ytitle(Pearson Correlation)/*
*/title(Pearson Correlation) caption(Ratio of Between/*
*/ Small Area Variation to Within Error Variation) legend(off)
graph export "E:\PhD\PhDpart3\Simulate\Pearson_edit.pdf", as(pdf) replace

/*Here this section of code reads in the multiple stata sheets
here the stata files SURVReg.do, Sigeta.do and SURVe2reg.do
were created by Haslett et al (2013)*/

clear all
*set memory 700m
*set matsize 200

*want to generate a new set of results
*Don't need to do for the first set of estimates
*here the data is being saved to $results\CensusLogExp_r*c_`iR'
*where *c is the factor that the cluster level variance has been
*multiplied by in $panal\RanDraw_edit.do
*here c=0, 0.5, 1, 1.5, 2, 2.5, 3, 4, 5
*here `iR' is 1-5 where this represents the 5 data sheets
*the bootstrap population data is saved to.
global results "E:\PhD\PhDpart3\Cambodia_stata\2019\Results"
forvalues iR=1/5   {
use $results\CensusExp_r`iR', clear
```

```
drop Yb1-Yb100
save $results\CensusLogExp_r*c_‘iR’, replace
}

clear all

global sraw "E:\PhD\PhDpart3\Cambodia_stata\Data\CSES2009\Raw"
global snew "E:\PhD\PhDpart3\Cambodia_stata\Data\CSES2009\Created"
global craw "E:\PhD\PhDpart3\Cambodia_stata\Data\Census2008"
global cnew "E:\PhD\PhDpart3\Cambodia_stata\Data\Census2008"
global panal "E:\PhD\PhDpart3\Cambodia_stata\2019\Analysis"
global results "E:\PhD\PhDpart3\Cambodia_stata\2019\Results"
global temp  "E:\PhD\PhDpart3\Cambodia_stata\2019\Temp"


run "$panal\SURVReg.do"
run "$panal\SigEta.do"
run "$panal\SURVe2reg.do"

*changing this to RanDraw_edit
*this creates 101 bootstrap estimates for each household in the population
*This can take several hours to run. The computational strength will
*of the computer determines how long it takes to run.
*For example on one of the computers used it took about 2.45 hours and on a
*laptop used it took about 5 hours to run.

global j=1
while $j<=101                          {
run "$panal\RanDraw_edit.do"
global j=$j+1
}

*$panal\SURVReg.do
*This model was created by Haslett et al (2013)
set more off
use $snew\CSES_povmodel.dta, clear

*Final model (????) :
*Reduce:
*this is the orginal model before it was started to be changed
#delimit;
global xvar "hhsize lnhhsz pkids06 plit pseced notoilet numroom rfree
car cellphone computer electric motorbike phone radio tv
floor_t floor_c floor_s roof_t roof_c roof_m wall_b
```

```
boat_e cellphone_e h_lit_e plit_e resplus_e reg3
hhsizeXS3 roof_cXS3 numroomXS3 motorbikeXS3
"; #delimit cr


svy: reg ln_exp $xvar


* Things to keep
predict r, resid
save "$snew\SURVEYr", replace
global survN=_N
global px=e(df_m)
matrix beta=e(b)'
matrix Vbh=cholesky(e(V))
do "$panal\SigEta"

*$panal\SigEta
*This data file was created by Haslett et al (2013)
/* Splits residuals into cluster-level hi in PSUerr.dta*/
/* and household level e in HHERR.dta                  */
/* Uses ELL method for estimating variance components  */
/* Requires results from fit and residuals in SURVEYr  */
*version 9.1

/*This is adapted code from SigEta from ELL
This is the ELL approximation to calucalte the variation between the
small areas*/

use "$snew\SURVEYr", clear
preserve
collapse (mean) hi=r (sum) w_c=pweight (count) nc=r, by(psu)
global survC=_N
gen Nc=sum(nc)
sort psu
save "$snew\PSUerr", replace
use "$snew\SURVEYr", clear
sort psu
merge psu using "$snew\PSUerr"
drop _merge
egen w_T=sum(pweight)
replace w_c=w_c/w_T
gen e=r-hi
save "$snew\HHerr", replace
```

```
gen e2=e*e
gen r2=_N*pweight*r*r/w_T
scalar dfd=e(N)-e(df_m)-1
collapse (mean) hi=hi w=w_c nc=nc (sum) r2=r2 e2=e2, by(psu)
gen tau2=e2/(nc*(nc-1))
gen num1=w*hi*hi
gen num2=w*(1-w)*tau2
gen denom=w*(1-w)
collapse (sum) num1=num1 num2=num2 denom=denom rss=r2
gen sig2eta=(num1-num2)/denom
gen sig2u=rss/dfd
list sig2eta sig2u
display sig2eta/sig2u
restore
drop r

*$panal\SURVe2reg.do
*created by Haslett et al (2013)
/* Error variance Regression of L(e2) on Z              */
/* Use stepwise first to identify appropriate $zvar     */
/* Saves coefficients in matrix alpha, variance in Va */
/* residual variance in $rvar and A in $A                 */
/* Household-level standardized residuals estar are  */
/* stored in HHerr1.dta for resampling later              */


use "$snew\HHerr", clear

gen esq=e*e
summ esq, meanonly
global A=1.05*r(max)
gen L=ln((0.0001+esq)/($A-esq))

/*
sw, pr(0.1) pe(0.05) forward: reg L $xvar [pw=pweight]
*  0.5% on 6 vars
*/

svyset [pweight=pweight],strata(stratum) psu(psu)

*Current model R2=0.002 on p=2 vars
global zvar "reg2 reg3"
svy: reg L $zvar
```

```
predict B
predict rs, resid
matrix alpha=e(b)'
matrix Vah=cholesky(e(V))
replace B=exp(B)
global pz=e(df_m)
gen rs2=rs*rs
summ rs2, meanonly
global s2r=r(mean)
gen sde=sqrt(($A*B-0.0001)/(1+B)+$s2r/2*($A+0.0001)*B*(1-B)/(1+B)^3)
gen estar=e/sde
sort psu
save "$snew\HHerr1", replace

collapse (mean) ebar=estar, by(psu)
sort psu
merge psu using "$snew\HHerr1"
drop _merge
replace estar=estar-ebar
keep psu estar
sort psu
save "$snew\HHerr1", replace

*$panal\RanDraw_edit.do
*This is adapted from Haslett et al (2013)
/* Prediction of Yb for census data                                    */
/* Predictions Yb1-Yb100 into $Rname.dta                               */
/* starting with census data in $CensusExp_vx.dta                     */

/* This version uses a single model, kept in SURVreg and SURVe2reg     */

/* version 8.2
set more off  */


/* Random beta  */
drop _all
global px1=$px+1
set obs $px1
gen z=invnorm(uniform())
mkmat z
matrix bstar=beta+Vbh*z

/* Random alpha  */
```

```
drop _all
global pz1=$pz+1
set obs $pz1
gen z=invnorm(uniform())
mkmat z
matrix astar=alpha+Vah*z

/* Loops done differently in Stata10 */

forvalues iR=1/5   {
use $cnew\CensusExp_v‘iR’, clear
global censC=psuc[_N]
gen xb=bstar[$px1,1]
forvalues i = 1/$px {
local v : word ‘i’ of $xvar
replace xb=xb+bstar[‘i’,1]*‘v’
}
gen za=astar[$pz1,1]
forvalues i = 1/$pz {
local v : word ‘i’ of $zvar
replace za=za+astar[‘i’,1]*‘v’
}
gen Bb=exp(za)
save "$temp\CENSUSw", replace  /*Working copy of census file*/

/* Sample psus in PSUerr*/
use "$snew\PSUerr", clear
if _N<$censC set obs $censC
gen rc=int(uniform()*$survC)+1
save "$temp\PSUerrb", replace

/* Now merge with HHerr1 and PSUerrb, to construct Y*/
use "$temp\CENSUSw", replace
merge using "$temp\PSUerrb"
gen psub=rc[psuc]
gen hb=hi[psub]
gen ncb=nc[psub]
gen Ncb=Nc[psub]
drop _merge

*Alison Note: multiplying hb by the factor of c as hb is the
*cluster level error
/* Draw estar from within cluster chosen for hi */
/* (hence peculiar messing about with ncb)      */
```

```
merge using "$snew\HHerr1"
gen rn=Ncb-int(uniform()*ncb)
gen estarb=estar[rn]
gen sdb=sqrt(($A*Bb-0.0001)/(1+Bb)+$s2r/2*($A+0.0001)*Bb*(1-Bb)/(1+Bb)^3)
gen eb=estarb*sdb
drop _merge
/*Here the logYb$j and Yb$j are adpated by multiplying the cluster
level error (hb) by a factor of c where c=0, 0.5, 1, 1.5, 2, 2.5,
3, 4, and 5 */
gen logYb$j=xb+hb+eb
gen Yb$j=exp(xb+hb+eb)
drop psuc-eb

*saving this to LogExp_r*c_‘iR’ where *c is the magnitude
*the error has been multiplied by e.g. 0, 0.5, 1, 1.5, 2, 2.5,
*3, 4, 5

sort ic
save "$temp\CENSUSw", replace
use "$results\CensusLogExp_r*c_‘iR’", clear
sort ic
merge ic using "$temp\CENSUSw"
drop _merge
save "$results\CensusLogExp_r*c_‘iR’", replace
}

*This uses the bootstrap estimates and creates the cluster level
*and small area estimates
*This is adapted from Haslett et al (2013)
*This data file is run *c times.
*The time taken to run will depend on the computational strength of
*the computer. Personally it took about 15 minutes for one
*of the computers used and it took close to an hour each time it
*was run on a laptop used

/* Creates small-area estimates from Yb1-Yb101                                    */
/* mcode = municipal code (regn + prov + mun)                                     */
/* Needs poverty levels defined by municipality and urbanity in PovLines.dta      */
/*                                                                                */
/* This version gives poverty incidence  at all levels (optionally urban/rural)   */
/* For other measures edit replacements of Yb‘i’                                  */
/*

global sraw "E:\PhD\PhDpart3\Cambodia_stata\Data\CSES2009\Raw"
```

```
global snew "E:\PhD\PhDpart3\Cambodia_stata\Data\CSES2009\Created"
global craw "E:\PhD\PhDpart3\Cambodia_stata\Data\Census2008"
global cnew "E:\PhD\PhDpart3\Cambodia_stata\Data\Census2008"
global panal "E:\PhD\PhDpart3\Cambodia_stata\2019\Analysis"
global results "E:\PhD\PhDpart3\Cambodia_stata\2019\Results"
global temp  "E:\PhD\PhDpart3\Cambodia_stata\2019\Temp"

*this is taking the small area level mean at expenditure and log expenditure
*level
*here in $results\CensusLogExp_r*c_‘iR’.dta *c represents
*the amount the cluster was multiplied by
forvalues iR=1/5   {
use $results\CensusLogExp_r*c_‘iR’.dta, clear
drop if Yb1==.
local i=1
while ‘i’<=101 {
gen Ybpov‘i’=(Yb‘i’<pline)
local i=‘i’+1
}
compress
collapse (count) size=Yb1 (mean) logYb1-Ybpov101 [pweight=hhsize], by(psuc) fast
sort psuc
merge psuc, using $cnew\CensusExp_areas‘iR’.dta
drop _merge
order province-rural reg2 reg3

if ‘iR’==1   {
save $results\P0ea_p_*c.dta, replace
}
if ‘iR’>1   {
append using $results\P0ea_p_*c.dta
save $results\P0ea_p_*c.dta, replace
}
}

compress
sort psuc
save $results\P0ea_p_*c.dta, replace

use $results\P0ea_p_*c.dta, clear
order province district commune village ea ezone rural reg2 reg3 psuc size /*
*/ Yb* logYb*
order province district commune village ea ezone rural reg2 reg3 psuc size /*
*/ Yb1-Yb101 logYb* Ybpov*
```

```
drop building hhno
sort province district commune


*Commune level:
collapse (count) Npp=Yb1 (mean) Yb1-Ybpov101  [pweight=size], by(province district commune) fas
order province district commune Npp Yb1-Yb101 logYb* Ybpov*
save $results\P0commune_*c.dta, replace

*This code is run *c number of times
/*It is used to generate \kappa and the Spearman and Pearson
correlation for the comparison of the ‘true’ small area statistics
and the mean of the 100 bootstrap estimates.
This is repeated for each of the 101 ‘true’ statistics and the
bootstrap estimates.
This process is then run *c times for each of the different
values that the cluster level variability was multiplied by
*/

use $results\P0commune_*c.dta, clear
order province district commune Npp Yb* logYb* Ybpov*
order province district commune Npp Yb1-Yb101 logYb* Ybpov*

egen Yb1e=rmean(Yb2-Yb101)
sum Yb1e
gen sdYb1e=r(sd)
egen SEYb1=rsd(Yb2-Yb101)
sum SEYb1
gen seYb1=r(mean)
gen ratioYb1=sdYb1e/seYb1

egen Yb2e=rmean(Yb1 Yb3-Yb101)
sum Yb2e
gen sdYb2e=r(sd)
egen SEYb2=rsd(Yb1 Yb3-Yb101)
sum SEYb2
gen seYb2=r(mean)
gen ratioYb2=sdYb2e/seYb2

*need to do with i=1 and i=2 and i=100 and i=101 seperately
forvalues i=3/99{
local j=‘i’+1
local h=‘i’-1
egen Yb‘i’e=rmean(Yb1-Yb‘h’ Yb‘j’- Yb101)
```

```
sum Yb‘i’e
gen sdYb‘i’e=r(sd)
egen SEYb‘i’=rsd(Yb1-Yb‘h’ Yb‘j’-Yb101)
sum SEYb‘i’
gen seYb‘i’=r(mean)
gen ratioYb‘i’=sdYb‘i’e/seYb‘i’
}

egen Yb100e=rmean(Yb1-Yb99 Yb101)
sum Yb100e
gen sdYb100e=r(sd)
egen SEYb100=rsd(Yb1-Yb99 Yb101)
sum SEYb100
gen seYb100=r(mean)
gen ratioYb100=sdYb100e/seYb100

egen Yb101e=rmean(Yb1-Yb100)
sum Yb101e
gen sdYb101e=r(sd)
egen SEYb101=rsd(Yb1- Yb100)
sum SEYb101
gen seYb101=r(mean)
gen ratioYb101=sdYb101e/seYb101


*Generaing the ratio for for log expenditure which is logYb

egen logYb1e=rmean(logYb2-logYb101)
sum logYb1e
gen sdlogYb1e=r(sd)
egen SElogYb1=rsd(logYb2-logYb101)
sum SElogYb1
gen selogYb1=r(mean)
gen ratiologYb1=sdlogYb1e/selogYb1

egen logYb2e=rmean(logYb1 logYb3-logYb101)
sum logYb2e
gen sdlogYb2e=r(sd)
egen SElogYb2=rsd(logYb1 logYb3-logYb101)
sum SElogYb2
gen selogYb2=r(mean)
gen ratiologYb2=sdlogYb2e/selogYb2

*need to do with i=1 and i=2 and i=100 and i=101 seperately
```

```
forvalues i=3/99{
local j=‘i’+1
local h=‘i’-1
egen logYb‘i’e=rmean(logYb1-logYb‘h’ logYb‘j’- logYb101)
sum logYb‘i’e
gen sdlogYb‘i’e=r(sd)
egen SElogYb‘i’=rsd(logYb1-logYb‘h’ logYb‘j’-logYb101)
sum SElogYb‘i’
gen selogYb‘i’=r(mean)
gen ratiologYb‘i’=sdlogYb‘i’e/selogYb‘i’
}

egen logYb100e=rmean(logYb1-logYb99 logYb101)
sum logYb100e
gen sdlogYb100e=r(sd)
egen SElogYb100=rsd(logYb1-logYb99 logYb101)
sum SElogYb100
gen selogYb100=r(mean)
gen ratiologYb100=sdlogYb100e/selogYb100

egen logYb101e=rmean(logYb1-logYb100)
sum logYb101e
gen sdlogYb101e=r(sd)
egen SElogYb101=rsd(logYb1- logYb100)
sum SElogYb101
gen selogYb101=r(mean)
gen ratiologYb101=sdlogYb101e/selogYb101

*Generaing the ratio for for poverty which is Ybpov

egen Ybpov1e=rmean(Ybpov2-Ybpov101)
sum Ybpov1e
gen sdYbpov1e=r(sd)
egen SEYbpov1=rsd(Ybpov2-Ybpov101)
sum SEYbpov1
gen seYbpov1=r(mean)
gen ratioYbpov1=sdYbpov1e/seYbpov1

egen Ybpov2e=rmean(Ybpov1 Ybpov3-Ybpov101)
sum Ybpov2e
gen sdYbpov2e=r(sd)
egen SEYbpov2=rsd(Ybpov1 Ybpov3-Ybpov101)
sum SEYbpov2
gen seYbpov2=r(mean)
```

```
gen ratioYbpov2=sdYbpov2e/seYbpov2

*need to do with i=1 and i=2 and i=100 and i=101 seperately
forvalues i=3/99{
local j=‘i’+1
local h=‘i’-1
egen Ybpov‘i’e=rmean(Ybpov1-Ybpov‘h’ Ybpov‘j’- Ybpov101)
sum Ybpov‘i’e
gen sdYbpov‘i’e=r(sd)
egen SEYbpov‘i’=rsd(Ybpov1-Ybpov‘h’ Ybpov‘j’-Ybpov101)
sum SEYbpov‘i’
gen seYbpov‘i’=r(mean)
gen ratioYbpov‘i’=sdYbpov‘i’e/seYbpov‘i’
}

egen Ybpov100e=rmean(Ybpov1-Ybpov99 Ybpov101)
sum Ybpov100e
gen sdYbpov100e=r(sd)
egen SEYbpov100=rsd(Ybpov1-Ybpov99 Ybpov101)
sum SEYbpov100
gen seYbpov100=r(mean)
gen ratioYbpov100=sdYbpov100e/seYbpov100

egen Ybpov101e=rmean(Ybpov1-Ybpov100)
sum Ybpov101e
gen sdYbpov101e=r(sd)
egen SEYbpov101=rsd(Ybpov1- Ybpov100)
sum SEYbpov101
gen seYbpov101=r(mean)
gen ratioYbpov101=sdYbpov101e/seYbpov101

*keep ratio*
*duplicates drop ratio*, force

*generate the correlation
*Spearman correlation
mat spearmatYb_*c= J(101,1,.)
mat spearmatlogYb_*c= J(101,1,.)
mat spearmatYbpov_*c= J(101,1,.)

forvalues i=1/101{
egen RYb‘i’=rank(Yb‘i’), unique
egen RlogYb‘i’=rank(logYb‘i’), unique
egen RYbpov‘i’=rank(Ybpov‘i’), unique
```

```
egen RYb‘i’e = rank(Yb‘i’e), unique
egen RlogYb‘i’e = rank(logYb‘i’e), unique
egen RYbpov‘i’e = rank(Ybpov‘i’e), unique
}

forvalues i=1/101{
corr RYb‘i’ RYb‘i’e
scalar t = r(rho)
matrix spearmatYb_*c[‘i’,1]=t
}

forvalues i=1/101{
corr RlogYb‘i’ RlogYb‘i’e
scalar t = r(rho)
matrix spearmatlogYb_*c[‘i’,1]=t
}

forvalues i=1/101{
corr RYbpov‘i’ RYbpov‘i’e
scalar t = r(rho)
matrix spearmatYbpov_*c[‘i’,1]=t
}

svmat spearmatYb_*c
svmat spearmatlogYb_*c
svmat spearmatYbpov_*c

mat corrmatYb_*c= J(101,1,.)
mat corrmatlogYb_*c= J(101,1,.)
mat corrmatYbpov_*c= J(101,1,.)

forvalues i=1/101{
corr Yb‘i’ Yb‘i’e
scalar t = r(rho)
matrix corrmatYb_*c[‘i’,1]=t
}

forvalues i=1/101{
corr logYb‘i’ logYb‘i’e
scalar t = r(rho)
matrix corrmatlogYb_*c[‘i’,1]=t
}

forvalues i=1/101{
```

```
corr Ybpov‘i’ Ybpov‘i’e
scalar t = r(rho)
matrix corrmatYbpov_*c[‘i’,1]=t
}

svmat corrmatYb_*c
svmat corrmatlogYb_*c
svmat corrmatYbpov_*c

save $results\commune_*c_all, replace

drop province-RYbpov101e
drop if spearmatYbpov_*c==.
gen id = _n

save $results\corr_*c, replace

use $results\commune_*c_all, clear
drop spearmatYb_*c-corrmatYbpov_*c

*this keeps the ratios but does not keep the correlations
*at this stage
keep ratio*
duplicates drop ratio*, force
mkmat ratioYb1-ratioYb101, matrix(ratio_Yb2)
mkmat ratiologYb1-ratiologYb101, matrix(ratio_logYb2)
mkmat ratioYbpov1-ratioYbpov101, matrix(ratio_Ybpov2)
mat ratioYb_*c=ratio_Yb2’
mat ratiologYb_*c=ratio_logYb2’
mat ratioYbpov_*c=ratio_Ybpov2’
drop ratioYb1-ratioYbpov101
svmat ratioYb_*c
svmat ratiologYb_*c
svmat ratioYbpov_*c
rename ratioYb_*c1 ratioYb_*c
rename ratiologYb_*c1 ratiologYb_*c
rename ratioYbpov_*c1 ratioYbpov_*c
gen id=_n
mat drop ratio_Yb2 ratio_logYb2 ratio_Ybpov2
save "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_*c", replace

*This combines the data files together and adapting the names
*and order of the variables into a useable state. It then
*stacks the different data sheets together.
```

```
use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_0, clear
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_05
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_1
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_15
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_2
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_25
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_3
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_4
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_5
drop _merge
rename spearmat* spear*
rename corrmat* corr*

save E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_0, clear
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_05
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_1
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_15
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_2
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_25
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_3
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_4
drop _merge
merge 1:1 id using E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_5
drop _merge

save E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_all, replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
```

```
rename spearYb_01 spearYb_0
rename spearYbpov_01 spearYbpov_0
rename spearlogYb_01 spearlogYb_0
rename corrYb_01 corrYb_0
rename corrYbpov_01 corrYbpov_0
rename corrlogYb_01 corrlogYb_0

rename spearYb_051 spearYb_05
rename spearYbpov_051 spearYbpov_05
rename spearlogYb_051 spearlogYb_05
rename corrYb_051 corrYb_05
rename corrYbpov_051 corrYbpov_05
rename corrlogYb_051 corrlogYb_05

rename spearYb_11 spearYb_1
rename spearYbpov_11 spearYbpov_1
rename spearlogYb_11 spearlogYb_1
rename corrYb_11 corrYb_1
rename corrYbpov_11 corrYbpov_1
rename corrlogYb_11 corrlogYb_1

rename spearYb_151 spearYb_15
rename spearYbpov_151 spearYbpov_15
rename spearlogYb_151 spearlogYb_15
rename corrYb_151 corrYb_15
rename corrYbpov_151 corrYbpov_15
rename corrlogYb_151 corrlogYb_15

rename spearYb_21 spearYb_2
rename spearYbpov_21 spearYbpov_2
rename spearlogYb_21 spearlogYb_2
rename corrYb_21 corrYb_2
rename corrYbpov_21 corrYbpov_2
rename corrlogYb_21 corrlogYb_2

rename spearYb_251 spearYb_25
rename spearYbpov_251 spearYbpov_25
rename spearlogYb_251 spearlogYb_25
rename corrYb_251 corrYb_25
rename corrYbpov_251 corrYbpov_25
rename corrlogYb_251 corrlogYb_25

rename spearYb_31 spearYb_3
```

```
rename spearYbpov_31 spearYbpov_3
rename spearlogYb_31 spearlogYb_3
rename corrYb_31 corrYb_3
rename corrYbpov_31 corrYbpov_3
rename corrlogYb_31 corrlogYb_3

rename spearYb_41 spearYb_4
rename spearYbpov_41 spearYbpov_4
rename spearlogYb_41 spearlogYb_4
rename corrYb_41 corrYb_4
rename corrYbpov_41 corrYbpov_4
rename corrlogYb_41 corrlogYb_4

rename spearYb_51 spearYb_5
rename spearYbpov_51 spearYbpov_5
rename spearlogYb_51 spearlogYb_5
rename corrYb_51 corrYb_5
rename corrYbpov_51 corrYbpov_5
rename corrlogYb_51 corrlogYb_5

order spearYb* spearYbpov* spearlogYb* corrYb* corrYbpov* corrlogYb*
summarize spearYb_*  spearYbpov_*   spearlogYb_*  corrYb_*  corrYbpov_*  corrlogYb_*

save E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all,
order id spearYb_* spearlogYb_* spearYbpov* corrYb_* corrlogYb_* corrYbpov*
save E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
stack spearYb_0-spearYb_5, into(spearYb_all) clear
save "$temp\spear_Yb", replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
stack spearYbpov_0-spearYbpov_5, into(spearYbpov_all) clear
save "$temp\spear_Ybpov", replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
stack spearlogYb_0-spearlogYb_5, into(spearlogYb_all) clear
save "$temp\spear_logYb", replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
stack corrYb_0-corrYb_5, into(corrYb_all) clear
save "$temp\corr_Yb", replace
```

```
use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
stack corrYbpov_0-corrYbpov_5, into(corrYbpov_all) clear
save "$temp\corr_Ybpov", replace

use E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_all, clear
stack corrlogYb_0-corrlogYb_5, into(corrlogYb_all) clear
save "$temp\corr_logYb", replace

*stacking the ratio data
use  "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_all", clear
order id ratioYb_* ratiologYb_* ratioYbpov_*
save "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_all", replace
stack ratioYb_0-ratioYb_5, into(ratio_Yb_all) clear
save "$temp\ratio_Yb", replace

use  "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_all", clear
stack ratioYbpov_0-ratioYbpov_5, into(ratio_Ybpov_all) clear
save "$temp\ratio_Ybpov", replace

use  "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_all", clear
stack ratiologYb_0-ratiologYb_5, into(ratio_logYb_all) clear
save "$temp\ratio_logYb", replace

*combinging the data
use "$temp\ratio_Yb", clear
merge m:m _stack using "$temp\ratio_Ybpov"
drop _merge
merge m:m _stack using "$temp\ratio_logYb"
drop _merge
merge m:m _stack using "$temp\spear_Yb"
drop _merge
merge m:m _stack using "$temp\spear_Ybpov"
drop _merge
merge m:m _stack using "$temp\spear_logYb"
drop _merge
merge m:m _stack using "$temp\corr_Yb"
drop _merge
merge m:m _stack using "$temp\corr_Ybpov"
drop _merge
merge m:m _stack using "$temp\corr_logYb"
drop _merge

save "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_corr", replace
```

```
*This is producing figures 7.4-7.9
use "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\ratio_corr", clear

twoway (scatter spearYb_all ratio_Yb_all, mcolor(black) msize(small)/*
*/ msymbol(smcircle_hollow)), ytitle(Spearman Correlation)/*
*/ xtitle(Ratio of Standard Deviation to Standard Error) title(Spearman /*
*/Correlation for Expenditure) legend(off)
graph export "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\spear_Yb.pdf",/*
*/as(pdf) replace


twoway (scatter spearYbpov_all ratio_Ybpov_all, mcolor(black) msize(small)/*
*/ msymbol(smcircle_hollow)), ytitle(Spearman Correlation)/*
*/ xtitle(Ratio of Standard Deviation to Standard Error) title(Spearman /*
*/Correlation for Poverty) legend(off)
graph export "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\spear_Ybpov.pdf",/*
*/as(pdf) replace

twoway (scatter spearlogYb_all ratio_logYb_all, mcolor(black) msize(small)/*
*/ msymbol(smcircle_hollow)), ytitle(Spearman Correlation/*
*/ for log Expenditure) xtitle(Ratio of Standard Deviation to Standard Error)/*
*/title(Spearman Correlation for Log Expenditure) legend(off)
graph export "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\spear_logYb.pdf",/*
*/as(pdf) replace

twoway (scatter corrYb_all ratio_Yb_all, mcolor(black) msize(small)/*
*/ msymbol(smcircle_hollow)), ytitle(Peasron Correlation)/*
*/ xtitle(Ratio of Standard Deviation to Standard Error) title(Pearson /*
*/Correlation for Expenditure) legend(off)
graph export "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_Yb.pdf",/*
*/as(pdf) replace

twoway (scatter corrYbpov_all ratio_Ybpov_all, mcolor(black) msize(small)/*
*/ msymbol(smcircle_hollow)), ytitle(Pearson Correlation)/*
*/ xtitle(Ratio of Standard Deviation to Standard Error) title(Pearson /*
*/Correlation for Poverty) legend(off)
graph export "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_Ybpov.pdf",/*
*/as(pdf) replace

twoway (scatter corrlogYb_all ratio_logYb_all, mcolor(black) msize(small)/*
*/ msymbol(smcircle_hollow)), ytitle(Pearson Correlation/*
*/ for log Expenditure) xtitle(Ratio of Standard Deviation to Standard Error)/*
*/title(Pearson Correlation for Log Expenditure) legend(off)
```

```
graph export "E:\PhD\PhDpart3\Cambodia_stata\2019\Results\corr_logYb.pdf",/*
*/as(pdf) replace
```