2004

Interviewer effects in household surveys: estimation and design

Nicholas Darby Von Sanden

University of Wollongong

Interviewer Effects in Household Surveys: Estimation and Design

A thesis submitted in fulfilment of the requirements for the award of the degree

Doctor of Philosophy

from

University of Wollongong

by

Nicholas Darby von Sanden B.Sc.(Hons), B.Econ. ANU

School of Mathematics and Applied Statistics

2004

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

Nicholas Darby von Sanden

16 December, 2004

Acknowledgements

This project would never have come to fruition without the generous aid and support of my family, friends and colleagues. In particular I would like to thank my supervisors Professor David Steel and Associate Professor Yan-Xia Lin for their insight and dedication.

I am very grateful for the ongoing support of the Australian Bureau of Statistics, especially Bill, Frank, Geoff and everyone who helped push through the CURF in time to test my results. Thanks also to Craig, Mark and the entire Time Series Analysis section for making me feel welcome whenever I came to visit.

Special mention goes to the gang at UoW who helped make my time outside work so much fun. Cheers to Alex, Eric, Grant, Matt, Riccardo, Rob, Russell and Sandy. I learnt alot from you too! Sadly there hasn't been enough space to fit Russell's insights on beer and basketball, Rob's observations regarding Star Wars and Matt's soccer tips into this thesis. However, I would like to thank Grant for showing me how to draw a helicopter in LaTeX. For those not in Wollongong; Anthony, Britt, David, Fiona, James, Makoto, Nikos, Rado, Richard, Simon, Will and Wongas, thanks also for all your support!

Finally to Mum, Dad and Chwee I couldn't have completed this without you!

Abstract

Although interviewers play an important role in the collection of high quality data, the presence of the interviewer may result in an unintended correlation between responses in surveys. The interviewer effect can greatly increase the variance of results derived from surveys and is not considered in standard variance estimates. The impact of interviewers on total survey error should be estimated so that survey data can be interpreted appropriately.

Previous studies examining the interviewer effect have generally relied on fully interpenetrated designs, in which a minimum of two interviewers are randomly allocated to each spatial area. Interpenetration provides repeated measurement of spatial areas which allows spatial and interviewer effects to be disentangled. However, conducting an interpenetrated survey is an expensive process that is not often applied in practice.

This thesis presents a review of techniques for estimating the interviewer component of total survey error and establishes a general framework for examining the interviewer effect in household surveys. The framework is applied to explicitly consider the relationship between the survey design and estimation of the interviewer effect for the first time. We demonstrate how the interviewer effect can be estimated in a variety of new scenarios, including designs in which the interviewer and spatial effects were previously considered confounded. We then consider the potential gain from incorporating the longitudinal and spatial information available to the survey designer.

The concept of partial interpenetration is introduced and clearly defined. Methods for estimating the interviewer effect in partially interpenetrated surveys are then considered. Partially interpenetrated survey designs will now allow us to estimate the interviewer effect and its contribution to total survey error more efficiently and cost-effectively as a regular part of the survey process. Procedures for preparing cost-optimal partially interpenetrated survey designs for the estimation of the interviewer effect are introduced and the gains from estimating the interviewer effect in

partially interpenetrated surveys quantified.

Regular estimation of the interviewer effect will result in higher quality surveys and will have positive implications for ongoing monitoring leading to more appropriate interviewer training, questionnaire design and cost-effective surveys.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Surveying is routinely applied by governments to gather information for determining social and economic policy. Given the size and expense involved in conducting these surveys it is important that they are constructed as accurately and efficiently as possible.

In many surveys information is collected directly from people within a household via a personal interview. The role of the interviewer varies across surveys. Interviewers may elicit participation, collect or edit data and prompt response. Well trained interviewers can have a positive effect on data quality, however the interviewer is also a source of error and variability in household surveys. Estimating and controlling for the impact of the interviewer effect is an important step in preparing a high quality survey and providing users with an indication of the level of certainty they can associate with survey data.

In practice estimates of the interviewer effect are generally not provided for data derived from household surveys. This occurs because estimation of the interviewer effect is problematic in area-based surveys where the interviewer effect cannot be fully separated from any spatial effects. Current techniques for isolating the interviewer effect have proven expensive and are rarely applied in practice.

This thesis will present a major extension to the range of survey designs from

which the interviewer effect can be estimated and demonstrate how the interviewer effect can be routinely estimated in practice. Improvements to the cost-effectiveness of survey designs will facilitate more regular estimation of the interviewer effect in practice, leading to improvements in survey design and interviewer training and more appropriate usage of data derived from area-based household surveys.

1.1 What Constitutes a Survey?

Surveying refers to a general act of gathering information. The survey methodologist, designing a complex, large-scale, survey for the purpose of making inferential statements about a *population* would define the term survey in more specific terms. In general there are two main types of survey

- the complete enumeration survey, or census. In this case the entire population is surveyed.
- the sample, or probability, survey. In this case a sub-sample of the relevant population is included.

Although the census is an important tool, the majority of modern surveys are sample surveys, which can be applied to make quantitative estimates of population characteristics.

According to O'Muircheartaigh (1997) modern surveys have developed from mostly independent research compiled in three main areas: governmental/official statistics; academic/social research; and commercial/advertising/market research. Due to the varied backgrounds of survey practitioners and the diverse range of situations to which surveys are applied, Kruskal (1991, p xxvii) was unable to comprehensively define a survey via its characteristics, stating;

'*That leaves me without sharp boundaries for this subtaxonomy of statistics, a condition of mild discomfort yet honesty.*'

As a consequence it is generally easier to either describe surveys with reference to a specific area of interest, or to define surveys according to their general aims.

Following an extensive review, Lessler and Kalsbeek (1992, p 1) attempt to comprehensively define a survey;

'*A survey is a scientific study of an existing population of units typified by persons, institutions or physical objects. A survey attempts to acquire knowledge by observing the population as it naturally exists and making quantitative statements about aggregate population characteristics.*'

This broad definition fails to encompass several types of survey such as modern market research targeting interest groups. In these surveys there may only be an interest in gathering qualitative information, perhaps for the purpose of detecting trends in fashion or music. This definition is, however, appropriate to the task of quantitatively examining non-sampling error in surveys.

Sudman *et al.* (1996, p 1) describe surveys in much more general terms;

'*We start with a dual conception of the survey. On the one hand, a survey is a social encounter. On the other hand, it is a series of cognitive tasks to be performed by respondents.*'

Sudman *et al.* (1996) do not try to provide a comprehensive definition of a survey, but their description shows a marked philosophical departure from Lessler and Kalsbeek (1992). This is in part due to the aim of Sudman *et al.* (1996); to provide a conceptual framework for examining the survey process utilizing a psychological perspective.

It is the variety of philosophical and practical differences that exist between survey practitioners that makes it difficult to pin down the characteristics of a survey. Perhaps the technique of introducing sample surveys by way of example, such as polling voters in New York City, (Yamane, 1967), is the most appropriate. In this way surveys can be simply and cogently defined relevant to a particular topic of interest.

For the remainder of this thesis we will be examining area-based, household surveys. An example is the Labour Force Survey (LFS) conducted by the Australian Bureau of Statistics (ABS). The LFS is a large multistage panel survey in which approximately 30,000 households are interviewed each month by asking any respon-

sible adult to supply details for both themselves and the remaining members of their household. Each household is generally enumerated by the same interviewer for eight consecutive months before being rotated out of sample along with one-eighth of the remaining units. The first month in the survey is enumerated via personal interview whilst the remainder are enumerated, where agreed, over the telephone.

The LFS is the primary source of labour statistics in Australia. Movements in employment and unemployment statistics compiled from this survey are scrutinized closely by policy makers, social welfare groups and the media. Consequently it is important that measures indicating the quality of these statistics are prepared to give users a clear indication of the reliability of this data.

1.1.1 Errors in Surveys

In general, if we consider a single survey producing some hypothetical set of descriptive summary statistics, the criteria which could be used to consider the quality of such a survey would depend on the aims of the users. For example some users might be interested in fast, approximate results and consider the timeliness of the survey to be of paramount importance. Other users might be more interested in very accurate results, regardless of the cost and length of time collecting this data would require. Thus different sets of users may have different and potentially conflicting criteria upon which they judge the quality of results collected in a survey.

Lyberg *et al.* (1997a, p xiii) recognize this conflict;

'*Indeed, some survey quality goals are conflicting, which introduces an irrational element into the decision making.*'

Given these potentially conflicting aims, it is important to have an understanding of the various quality goals relevant to each survey. A list of these goals has been compiled in the quality frameworks of national statistical agencies such as Statistics Canada and the ABS (e.g. Brackstone, 1999). Once these frameworks have been defined we need quantitative measures describing how well these quality goals have been achieved. Lyberg *et al.* (1997b) focus on survey errors as one such quantifiable

measure. However, for survey errors to be quantifiable they must also be defined and the definition of a survey error is a topic of considerable debate in the survey literature. O'Muircheartaigh (1997, p 1) recognizes that when defining the term error we encounter many of the same problems faced when defining the term quality;

'*The concept of quality, and indeed the concept of error, can only be defined satisfactorily in the same context as that in which the work is conducted. To the extent that the context varies, and the objectives vary, the meaning of the error will also vary.*'

O'Muircheartaigh (1997) proposes a very general definition of the term error, suggesting that errors occur when a survey purports to do what it does not do.

O'Muircheartaigh's definition can be seen as a general form of a more common definition adopted by Thompson (1997, p 6) amongst others;

'*The total survey error is the amount by which the estimate differs from the true value of the quantity for the target population.*'

Both of the above definitions agree that errors occur when a survey fails to achieve its purpose and produces an estimate other than an ideal result based upon the intentions of the user. However, these definitions can diverge conceptually when the nature of reality, or the true value, is considered.

1.1.2 The Concept of a True Value

When conducting a household survey we can consider the true value for an individual respondent to be a perfect recording of the information that we were interested in collecting from that individual. There are two schools of thought as to how this may occur in practice.

The first approach was initiated by the work of Hansen *et al.* (1951) and Sukhatme and Seth (1952). They believed that true values exist separately from the conditions under which the survey was run. Consequently over repeated trials of the survey (assuming the survey can be repeated infinitely with no memory of previous trials) the mean of the estimator will approach the true value. In this case the true value

exists as a characteristic of both the set of all samples that were possible to select and also the population. Repeated trials of the survey could then be used in an attempt to measure this characteristic. This approach has had a strong influence on later work, in particular Felligi (1964, 1974); Kish (1965); Raj (1968); Moser and Kalton (1972); Lessler (1977, 1983) and O'Muircheartaigh (1977).

Later authors criticized this approach, suggesting that the requirement of survey repeatability was extreme. Some authors, notably Deming (1950, 1960); Kish (1962) and Zarkovich (1966) argued that in practice it would be impossible to successfully repeat a trial of a survey, because of the myriad conditions that fluctuate upon repetition and also memory build-up associated with its running. As a consequence they did not believe that objective true values exist and instead suggested that true values could only be defined with reference to the conditions under which the survey was run, which they termed the *essential survey conditions*.

Hansen *et al.* (1967) acknowledged this point of view and the difficulties associated with defining operationally independent true values by introducing a concept, which could be described as a *preferred true value*. The rationale behind this concept was that in practice it might not be feasible to run a survey to measure exactly a particular concept and if this were the case a perfectly conducted survey would still not measure the objective true value. The concept of the preferred true value was therefore developed to reflect the ideal results that would be obtained if the survey were to be conducted perfectly under a practically ideal methodology, even if that methodology were not theoretically ideal.

Conceptually the notion of the preferred true value, operationally dependent on the essential survey conditions, can now be adopted by both sides of this debate. In contrast, defining an objective true value independent of the survey conditions remains a choice adopted by the individual methodologist. In practice authors such as O'Muircheartaigh (1997) have been able to skirt this debate with a more general definition of error based upon the aims of a survey. Under O'Muircheartaigh's (1997) definition if the survey purports to measure an objective true value that exists independent of the survey conditions then the survey error can be defined

with regards to these terms. On the other hand if it were not feasible to construct a survey to measure a desired concept but only to measure a preferred proxy, the survey error can be defined according to this preferred true value. Consequently O'Muircheartaigh's (1997) definition can be used to span both sides of this debate and this approach will be adopted in the remainder of this thesis when we consider survey errors such as the interviewer effect.

1.2 Total Survey Error

Errors associated with conducting sample surveys can be split into two main groups; sampling errors and all other sources of error; the non-sampling errors. Generally only the sampling error is estimated in practice and consequently the term total survey error is used to distinguish models that consider *both* sampling and non-sampling errors, such as the interviewer effect.

1.2.1 Sampling Errors

Sampling error arises due to differences between the characteristics of our sample and the population in which we are interested. We can use the sampled elements to make inferential statements regarding the entire population. However, if the sample were to have different characteristics from the rest of the population then our statements would not reflect the characteristics of the population, resulting in an error due to sampling. Groves (1989) refers to sampling error as an error of non-observation caused by heterogeneity of units in the population. Thus in the case of a census where we are able to observe the entire population, the sampling error will be reduced to zero. All other sources of error associated with the process of surveying can be described as non-sampling errors.

Neyman (1934) provided a framework through which sampling error could be examined. Subsequent authors have elaborated upon this framework and have established comprehensive theories regarding the estimation and treatment of sampling

errors. Classic texts in this field such as Deming (1950); Cochran (1977) and Kish (1965) elaborate upon sampling error theory.

1.2.2 Non-sampling Errors

Along with the development of sampling theory, there has also been extensive interest in non-sampling errors. A wide body of work dealing with one or more source of non-sampling error has been established. Despite occasional success many sources of non-sampling error have proven unresponsive to suggested treatments and have also proven hard to estimate or even define. Lyberg *et al.* (1997b) recognize two reasons for this; firstly non-sampling errors are usually not additive, i.e. a reduction in one source of non-sampling error may or may not lead to an increase in another. Secondly these errors are caused by a large number of potentially uncontrollable factors, further obscured by complex survey designs. Lyberg *et al.* (1997a, p xiii) recognize this as a weakness in the sample survey literature stating;

'*Comprehensive theories exist for the treatment of sampling errors. As for non-sampling errors, no such theory exists.*'

Due to the complexities inherent in estimating non-sampling errors and the relative ease of estimating sampling errors, sampling error estimates are often the only error estimate made available upon the completion of a survey. Groves (1989, p 293) realizes that this may lead to misleading estimates of the error, stating;

'*Sampling error represents only one component of total error. Treating standard errors as the reflection of total error ignores biases in statistics that can arise from coverage, non-response and measurement errors. In addition, such a view ignores the variability in survey results due to these same error sources.*'

Although, it may be reasonable to assume that non-sampling errors have a minor impact compared to sampling errors in some surveys, if they are ignored the survey results may still be rendered useless for a variety of purposes. Furthermore in many surveys non-sampling errors may have a greater impact than the sampling error. At the very least, confidence bounds around statistics based upon this assumption will

lead to potentially erroneous results and inappropriate usage. A necessary first step in the process of reducing non-sampling errors, such as the interviewer effect, is to define and estimate the impact of these errors.

The Definition and Estimation of Non-sampling Errors

The term non-sampling error refers to all sources of error, excluding the sampling error, that are associated with running a survey. Given that Groves (1989) considers sampling error to be an error of non-observation it is perhaps complementary to view non-sampling errors as errors that occur due to imperfections in both the process and act of observing our sample (although it must be noted that Groves describes observation errors independently). Under this notion non-sampling errors would completely disappear under conceptually ideal circumstances. Unfortunately, non-sampling errors are rife in even the most carefully designed survey, due to practical problems such as defining a frame, wording questionnaires and collecting data. Although a variety of methods exist for controlling non-sampling errors, it is difficult to evaluate the success or failure of these methods if we are unable to estimate the size of these sources of non-sampling error during the practical application of a survey.

There have been a number of attempts to create a comprehensive classification of non-sampling errors and a framework for their evaluation. Early classifications suggested by Bowley (1915, 1926) were expanded by Deming (1944) to a list of more than a dozen '*factors which affect the ultimate usefulness of a survey*'. Later authors began to limit the length of this list, describing more general areas of non-sampling error and developing a hierarchical classification structure. Cochran (1977, p 359) recognized three general areas of non-sampling error;

1. *'Failure to measure some of the units in the chosen sample. This may occur by oversight, or, with human populations, because of failure to locate some*

individuals or their refusal to answer the questions when located.

2. *Errors of measurement on a unit. The measuring device may be biased or imprecise. With human populations the respondents may not possess accurate information or they may give biased answers.*

3. *Errors introduced in editing, coding and tabulating the results.'*

Subsequent authors have classified errors according to where they occur in the survey process (Zarkovich, 1966), the techniques that are required for their estimation (Deming, 1960), the types of activities associated with the survey process (Lessler and Kalsbeek, 1992) and the contribution they make to the mean square error model (Kish, 1965; Groves, 1989).

Due to these different methods of classification, there is a certain amount of overlap in the various definitions applied to subsets of non-sampling error. For example, measurement error as defined by Beimer *et al.* (1991) is very similar to that of Cochran (1977). Lessler and Kalsbeek (1992) include both errors of measurement and errors introduced during the coding and imputation of results in their definition of measurement error. Even the terminology applied is not consistent as the measurement error defined by Cochran (1977) is broadly consistent with the notion of response error as defined by Sudman and Bradburn (1974).

Measurement Error

Following Sarndal *et al.* (1992), this thesis will concentrate upon measurement error as a subset of all non-sampling errors. Measurement error is described by Groves (1989) as an error that occurs during the act of observation, where for whatever reason the value required by the survey design is not recorded correctly. Once we have decided upon a survey design and instrument and begun surveying an element, measurement error refers to any inappropriate observations made regarding this element. In contrast other non-sampling errors, such as non-response and coverage

error, relate to the process of appropriately obtaining the units in the sample before the act of observation.

There are four main sources of measurement error in a household survey; the respondent, interviewer, instrument (e.g. questionnaire) and the mode of collection (e.g. over the telephone). Each of these sources has potential to cause a miscommunication of the intent of a survey question to the respondent and lead to an incorrectly recorded or observed response. The focus of this thesis will be to consider how we can estimate the contribution of the interviewer to the measurement error.

1.2.3 The Interviewer as a Source of Error

The interviewer effect can be considered to be the component of measurement error that can be attributed to the presence and characteristics of the interviewer conducting a survey. The act of surveying itself can be viewed as a complex process of social interaction and consequently it is difficult to conceive of separately isolable contributions from both the interviewer and the respondent to the measurement error. Difficulties with isolating an interviewer effect have been recognized by Sudman and Bradburn (1974) who consider interviewer main effects to be meaningless and O'Muircheartaigh (1997, p 12) who states;

'*It is impossible to separate entirely the function and behaviour of the interviewer from the function and behaviour of the respondent in social surveys.*'

We recognize, however, that some interviewers have a systematic impact on the responses that they collect. Data collected by these interviewers will generally be more similar than it would have been if collected by different interviewers. Thus the presence of the interviewer results in an unintended correlation in survey data. This correlation has a negative impact on results derived from surveys and has proven difficult to eradicate. We can therefore define the interviewer effect we wish to measure as the negative impact on the precision of survey estimates due to the correlation of responses caused by the presence and characteristics of the interviewer.

Estimation of this correlation will allow us to cater for the interviewer effect and make appropriate estimates in household surveys.

Bassi and Fabbris (1997, p 733) recognize that this definition of the interviewer effect covers both a single interviewer's bias and and the interviewer effect as a variance component of the total survey error.

'*Interviewer error may be related to the interviewer's personal characteristics, attitudes, and behaviour during the interview. The error is a bias if it is referred to a single interviewer. On the other hand, it is a correlated random error if it is referred to the sample.*'

Consequently our definition of the interviewer effect covers both interviewer bias, if we consider the impact on response of a single interviewer, and the overall interviewer effect on total survey error.

Current Research

The interviewer effect was recognized in the early social surveys of the 20th century. Rice (1929) realized that interviewers with different political opinions tended to obtain different results in a survey of destitute men, while Mahalanobis (1946) and Hochstim and Stock (1951) attempted to estimate interviewer effects. These early studies concentrated on establishing the existence of a basic interviewer effect while later studies, such as Hansen and Marks (1958), attempted to establish the relative importance of the interviewer effect in comparison with other sources of error.

Subsequently two different models for investigating the interviewer effect were defined in the 1960s. Hansen *et al.* (1961) describe the interviewer effect as a correlation between responses collected by the same interviewer. This is commonly referred to as the US Bureau of the Census model. Kish (1962) proposed an alternative ANOVA or linear models approach which assumes that the interviewer effect can be described as an additive component in a linear model. The US Bureau of the Census model has influenced the work of many authors and was extended by Felligi (1964); Murthy (1967); Raj (1968); Bailar (1968); Bailar and Dalenius (1969); Koch (1973); Koop (1974) and Lessler (1976). Though less commonly applied, the

ANOVA approach was extended by Sukhatme and Sukhatme (1970); Hartley and Rao (1978); Biemer (1978) and Biemer and Stokes (1985).

The increased power of computing facilities, extended use of statistical models in data analysis and issues raised during the design-based and model-based debate amongst survey methodologists (see Chapter 2 for more detail) led O'Muirceartaigh and Wiggins (1981) to incorporate measurement errors directly into a statistical model for the examination of aircraft noise on annoyance. Their log-linear approach was innovative but has proven less influential than the multi-level modelling approach adopted by Anderson and Aitken (1985) who investigated interviewer variability in a survey on consumer spending. The multi-level modelling approach has the advantage that it can be used to cater for the hierarchical structure of datasets. Subsequent work by Hox *et al.* (1991); Pannekoek (1991); Wiggins *et al.* (1992); Hox (1994); Goldstein (1995); Pickery and Loosveldt (2000, 2001, 2004); O'Muircheartaigh and Campanelli (1998, 1999) and Martin and Beerten (2002) have applied and extended this approach.

These authors generally assume either an explicitly interpenetrated (see Chapter 4 for more detail) survey design in which a minimum of 2 interviewers are allocated to each region or an effectively interpenetrated design in which all of the interviewers are allocated to a single concentrated geographic region. This occurs even in the case of Pickery and Loosveldt (2000, 2001) who consider application of the longitudinal information available in repeated panel surveys. Interpenetration is assumed in order to avoid confounding between spatial and interviewer effects and allow estimation of the interviewer effect. Fully interpenetrated survey designs are generally considered too expensive to apply in practice and hence the interviewer effect is often ignored. This thesis will comprehensively define interpenetration for the first time and demonstrate how the interviewer effect can be estimated in cost-effective partially interpenetrated survey designs. We will also consider an extension to the work of Pickery and Loosveldt (2000, 2001) in which we show how repeated measurement of individuals over time by different interviewers can be considered as a form of longitudinal interpenetration.

It is now widely accepted that the characteristics of an individual interviewer will have a contribution to the overall measurement error in a survey. Although some authors, such as Felligi (1974) suggest self-completion questionnaires as a possible method to eradicate this source of error, the positive contribution of the interviewer in terms of overall data quality should not be overlooked. The role of the interviewer varies across surveys but as a minimum they are generally expected to elicit participation, collect/edit data and prompt response. Despite more than half a century of research into the effect of interviewer characteristics on survey response, many of the findings have proven inconsistent. For example Barr (1957); Freeman and Butler (1976) found that interviewer's attitudes affected response, while Feldman *et al.* (1951); Collins (1980) found the opposite. In summary Collins (1980) found that the only socio-demographic characteristic of interviewers to have been consistently found to affect survey response is the race of the interviewer.

Groves and Couper (1998) suggest that these inconsistent findings are caused by the interaction between the characteristics of both the respondent and the interviewer. This notion is also acknowledged by Sudman and Bradburn (1974) who considered it necessary to construct a comprehensive conceptual framework describing the entire interview process in order to examine response effects. Subsequent authors such as Groves *et al.* (1992); Schwarz and Sudman (1996); Campanelli *et al.* (1997) also believed that the effect of interviewers should only be considered in relation to all other factors which may affect response.

There is a large body of literature detailing the contribution of the characteristics of the interviewer to the interviewer effect, with particular focus on the effect of the interviewer on non-response and on modelling the entire interview process. Reviews can be found in Groves and Couper (1998); Sudman and Bradburn (1974); Groves *et al.* (1992); Schwarz and Sudman (1996); Groves *et al.* (2004); Beimer and Lyberg (2003) and Campanelli *et al.* (1997). In general the multilevel variance decomposition techniques developed in Chapters 3 and 6 of this thesis can be extended to consider non-response issues and the characteristics of the interviewer by incorporating appropriate covariates.

1.3 Interpenetrated Sampling and Cost Constraints

We have discussed how all previous studies for the estimation of the interviewer effect have relied on fully interpenetrated survey designs and that due to the high costs of interpenetration, interviewer effects are generally not estimated in practice. In this thesis we will focus on surveys conducted via personal interview, however interpenetration does not generally occur even for centralized telephone surveys. This thesis will provide the first comprehensive definition of interpenetration and introduce cost-effective, partially interpenetrated survey designs for the estimation of the interviewer effect in face-to-face surveys.

Interpenetration rarely occurs in large scale area-based surveys due to the costs associated with travelling between households. As large scale surveys are generally prepared under tight budgetary constraints it is common practice for interviewer allocations to be as geographically concentrated as possible. In remote areas, where travelling costs are higher, a single interviewer often enumerates an entire region and although it may be feasible to have two interviewers covering the same area in more densely populated cities, this rarely occurs in practice as it will lead to an increase in costs.

Generally individuals residing in a concentrated geographic area will share a number of common characteristics. For example the majority of people living in the most exclusive region of a city will likely have above average income. Thus if there is only one interviewer allocated to an entire geographical area, the correlation between responses introduced by the characteristics and actions of the interviewer will be confounded with the correlation between responses due to the similarity of respondents living in that area.

In order to isolate the interviewer effect from the confounding effects of the region, Mahalanobis (1946) designed the technique of interpenetrated samples. This technique requires a minimum of two interviewers to be randomly allocated to each concentrated geographic area. Interpenetration is recognized by Collins (1980) as a minimum requirement for the isolation and estimation of the interviewer effect, and

also by Groves (1989, p 360) who states;

'*All designs for estimating interviewer variance share the feature that different interviewers are assigned to equivalent respondent groups.*'

We have seen that estimation of the interviewer effect is an integral component of the total survey error. Consequently new cost-effective survey designs must be prepared to facilitate regular estimation of the interviewer effect in practice. This thesis will introduce new techniques that can be applied to routinely estimate the interviewer effect in non-fully interpenetrated surveys.

1.4 Issues to be Explored

Estimates of the interviewer effect can be used to produce more appropriate variance estimates, giving users a better idea of the reliability they can associate with results derived from surveys. For the survey designer, estimates of the interviewer effect can also be used to focus interviewer training techniques and highlight questionnaire design issues. Producing estimates of the interviewer effect will therefore lead to more appropriate and cost-effective household surveys.

Previous research relating to the estimation of interviewer effects has only considered fully interpenetrated survey designs. Full interpenetration is too expensive to regularly apply in practice and cost-effective techniques for the estimation of the interviewer effect will now be considered. This thesis will introduce partially interpenetrated survey designs for estimating the interviewer effect. The concept of partial interpenetration will be defined for the first time and new methods for the estimation of interviewer effects in surveys that are not fully interpenetrated will be developed. These techniques will be demonstrated on data drawn from the Australian Labour Force Survey.

In this thesis we will discuss a number of specific topics relating to the estimation of interviewer effects in partially interpenetrated surveys. These topics will be addressed in Chapters 3 to 6 respectively.

In Chapter 3 the first comprehensive demonstration detailing how the interviewer

effect and its contribution to total survey error can be estimated in a partially interpenetrated survey will be presented. The potential increase in the effective degree of interpenetration though incorporation of the longitudinal information available in repeated panel surveys will also be considered.

In Chapter 4 a new comprehensive definition of interpenetration, explicitly defining partial interpenetration and confounding, will be presented. Methods for determining the degree of interpenetration, given a survey design will be introduced and an algorithm to evaluate the degree of interpenetration will also be demonstrated.

Based on this definition of partial interpenetration techniques for determining optimal partially interpenetrated survey designs will be presented in Chapter 5. Estimation of the interviewer effect in partially interpenetrated survey designs, optimized with respect to a representative cost function will be considered. A further extension to multiple objective optimal survey designs, simultaneously minimizing either the sampling or total variance and the variance of estimates of the interviewer effect will also be introduced.

In Chapter 6 the application of spatial correlation models to partially interpenetrated survey data will be considered. This will be the first time that the spatial processes underlying the spatial effects in interpenetrated survey designs will be explicitly considered. In this way the remaining spatial information available to the survey designer can be incorporated into the estimation of the interviewer effect.

In summary this thesis presents the first major extension to the range of survey designs from which the interviewer effect can be estimated since Mahalanobis introduced interpenetration in 1946. The introduction of partial interpenetration allows estimation of the interviewer effect for surveys in which the interviewer and spatial effects would previously have been considered confounded. A number of practical demonstrations, incorporating spatial and longitudinal information will be presented and optimal partially interpenetrated designs for the estimation of the interviewer effect will be derived. In this way we aim to greatly enhance the range and cost-effectiveness of survey designs from which the interviewer effect can be estimated, leading to regular estimation of the interviewer effect and more appropriate usage

of data derived from large scale surveys.

Chapter 2

Models for Examining the Interviewer Effect

In the previous chapter we saw how the interviewer effect is related to the correlation between responses collected by each interviewer. Consequently adjusting for the interviewer effect generally requires an appropriate estimate of this correlation. This chapter will discuss how estimation of the interviewer effect is an integral part of the total survey error models that are required for making appropriate inference in household surveys. An appropriate notation for discussion of the interviewer effect will be developed and relevant results will be derived.

In this chapter we will demonstrate the impact of the interviewer effect on survey data and show how ignoring the interviewer effect leads to an underestimate in the variance of estimates derived from surveys. Previous studies have generally not considered the impact of the interviewer effect on the total variance making it difficult to determine the relevance of the interviewer effect in comparison to the total survey error. In this chapter we will consider the total variance under an interviewer effect model and estimates of the total variance, adjusted for the interviewer effect, will be introduced.

It will be shown how results under classical techniques for the estimation of the

interviewer effect, such as the ANOVA and the US Bureau of the Census correlational approach can also be derived under the broader class of Generalized Linear Mixed Models (GLMMs). The appropriateness of multilevel models (and hence GLMMs) for the estimation of the interviewer effect has been demonstrated by Anderson and Aitken (1985) and Hox (1994). The impact of a number of competing survey designs on the estimability of the interviewer effect will be discussed.

GLMMs are a very flexible class of models which naturally lend themselves to consideration of the interviewer effect. Improvements in computing techniques have made it possible to estimate GLMM parameters in increasingly complex scenarios and thus we can now consider estimating the interviewer effect in household surveys.

In this chapter a general framework for considering interviewer effects will be developed. This framework will be applied to demonstrate how classical techniques can be utilized to estimate the interviewer effect. A further extension to this framework will then be presented in Chapter 4, in which a broader definition of interpenetration, incorporating the concept of partial interpenetration for the first time, will be developed.

2.1 Total Survey Error Models

A comprehensive total survey error model describing the different errors associated with various survey operations can provide users with both a framework which can be used to appropriately measure the accuracy of survey estimates and also provide an idea as to the relative importance of the various error sources. The need for such an error model was identified and developed in work by Hansen *et al.* (1951, 1961, 1964). They formulated what has become known as the US Bureau of the Census Survey Error Model which provided a general framework upon which much of the subsequent research into this area has been based.

Forsman (1989) identified two directions in which subsequent research into survey error models has generally been pursued. One direction has focused on the development of theory for specific sources of non-sampling errors, such as interviewer and

non-response effects, while the other focuses on developing an integrated model for the treatment of survey errors.

The desire to formulate an integrated approach to total survey error has been motivated by interrelationships between sources of error. For example, the presence of an interviewer may introduce a correlated error (i.e. the interviewer effect we are trying to measure) while simultaneously helping to reduce respondent error through techniques such as prompting and clarification. Generally attempts at formulating comprehensive total survey error models have been unsuccessful. Platek and Gray (1983) derived intractable expressions when attempting to formulate a total survey error model. Lessler and Kalsbeek (1992) acknowledge the cost and complexity involved with formulating these models. More recently Platek and Sarndal (2001, p 15) recognized the difficulties associated with defining and estimating all of the various components of error;

'*At present it seems unreasonable to have a functional survey error model, one that would account for the major errors and yield estimates of the various components.*'

The majority of research into specific sources of survey error has been developing in isolation despite what several authors recognize as the need for collaborative efforts. Platek and Sarndal (2001) suggest that innovation in this area appears to have been inhibited by cost and practicality constraints. In the absence of a comprehensive total survey error model, research into specific sources of error in survey data will lead to the gradual accumulation of a systematic body of knowledge and provides the best hope for eventual estimation of total survey error.

The following sections will consider the design-based and model assisted approaches to estimating sampling error. These approaches can be combined with variance decomposition techniques to produce a total survey error estimate. An example demonstrating how to estimate the total variance for a simple random sampling without replacement survey with interviewers is presented.

2.1.1 Design-based and Model-based Approaches to Estimation in Sample Surveys

The design-based approach to estimation in sample surveys has also been referred to as probability sampling (e.g. Sarndal *et al.*, 1992), p-based sampling (e.g. Smith, 1994), the representative method (e.g. Kiaer, 1897) and randomization theory (e.g. Lohr, 1999). In general terms the design-based approach makes no distributional assumptions regarding a population of interest, but assumes a probability sample has been selected, that is, every unit in the population has a known, non-zero, chance of selection. More specifically, if we are given a fixed, unobserved population of interest, u, containing $i = 1, ..., N$ elements from which we draw a sample, s, with probability of selection, $p(s)$, the probability of selecting the sample can be determined by the sampling design and then we can define an indicator variable, U_i, to identify the n population elements that have been drawn in the sample

$$U_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

From this definition we can make statements regarding the finite population of interest by considering the expectation of the indicator variable with respect to all possible samples that can be selected, as described by the probability sampling distribution. For example, under a Simple Random Sampling Without Replacement (SRSWOR) survey design, the probability that each population element i is included in the sample, π_i, will be equal for all elements in u

$$\pi_i = \Pr\left(U_i = 1\right) = \frac{n}{N}$$

The expectation of U_i with respect to the sample design will be

$$\begin{aligned} E_p\left[U_i\right] &= \sum\nolimits_{k=0}^{1} k \Pr\left(U_i = k\right) \\ &= \frac{n}{N} \end{aligned}$$

Where $E_p[.]$ is the expectation with respect to the sample design. Similarly $E_p[U_i^2] = \frac{n}{N}$ and $E_p[U_i.U_{i'}] = \frac{n.(n-1)}{N.(N-1)}$. For example when considering the number raised estimator for the population mean, $\hat{\bar{y}} = \frac{1}{n}\sum_{i \in s} y_i$, the estimator will have expectation of

$$\begin{aligned} E_p\left[\hat{\bar{y}}\right] &= E_p\left[\frac{1}{n}\sum\nolimits_{i \in s} y_i\right] \\ &= \frac{1}{n}\sum\nolimits_{i \in u} y_i E_p[U_i] \\ &= \bar{y}_u \end{aligned}$$

The variance of the estimator can also be derived based on the sampling distribution determined by the indicator variable. Then for SRSWOR $Var_p\left(\hat{\bar{y}}\right) = E_p\left[\left(\hat{\bar{y}} - \bar{y}_u\right)^2\right] = (1-\frac{n}{N})\frac{S_Y^2}{n}$ where $S_Y^2 = \frac{1}{N-1}\sum_{i \in u}(y_i - \bar{y}_u)^2$ see for example Cochran (1977, p 23).

The Model Assisted Approach

The design-based approach does not naturally lend itself to estimation of other sources of non-sampling error as required for total survey error models, because as Smith (1994, p 8) points out

'*Models are essential for dealing with all forms of non-sampling error, including coverage errors, nonresponse, response errors and processing errors, and it is also accepted that models help in the choice of estimators, especially when covariates are available.*'

The model assisted approach allows us to evaluate sampling as well as measurement errors by considering the generation of the observed survey to be a two-stage process, in which both the sample design and the measurement process, described by an appropriate model, contribute to the stochastic nature of the survey.

A useful modification, applied in this thesis, occurs when these two stages are independent so that the sampling design is ignorable and the measurement error model is valid for the entire population. Pfeffermann *et al.* (1998, p 24) recognized this

'*When the selection probabilities are related to the values of the response variable even after conditioning on the model covariates, the sampling process becomes informative and the model holding for the sample data is different from the population model. Ignoring the sampling process in such cases may yield biased point estimators and distort the analysis.*'
He proposed a probability weighting procedure to incorporate the effect of informative sampling.

If the sampling design is non-informative an appropriate response model will be valid for both samples drawn from the finite population of interest and also for the finite population itself. A statistical model can then be employed to reflect the errors associated with measuring data on sample members. If Y_i is an operational *true* value, y_i is an *observed* value and ε_i is the measurement error for population element i as observed in a sample survey, we might assume the following simple response error model

$$y_i = Y_i + \varepsilon_i \tag{2.2}$$

where $E_m\left[\varepsilon_i | i \in s\right] = 0$ and $V_m\left[\varepsilon_i | i \in s\right] = \sigma_\varepsilon^2$, the subscript m here refers to expectations with respect to the model, conditional on the sample. The simple response error model (2.2) captures the notion that, while the *true* values are fixed for each element i the measurement process is stochastic and hence the observed value y_i is a random variable. More complex response error models, including the use of multilevel models may also be used and these will be considered in a later section.

Under the model assisted approach, expectations of estimators with respect to the sample design $E_p[.]$ and the response model $E_m[.]$ can be evaluated conditional on the measured population values and on the selected sample respectively. Therefore given a non-informative sample design and an estimator, $\hat{\psi}$, the unconditional expectation can be calculated as a two stage process either conditioning on the model or the selected sample first, i.e.

$$E\left[\hat{\psi}\right] = E_p\left[E_m\left[\hat{\psi}|s\right]\right]$$

which in the case of our previous number raised estimator would lead to an unbiased estimate of the population mean if $E_m[\varepsilon_i] = 0$

The unconditional variance of $\hat{\psi}$ can also be determined conditioning on each stage. We can either recognize that

$$\begin{aligned} V\left(\hat{\psi}\right) &= E_p\left[V_m\left(\hat{\psi}|s\right)\right] + V_p\left[E_m\left(\hat{\psi}|s\right)\right] \\ &= E_m\left[V_p\left(\hat{\psi}|m\right)\right] + V_m\left[E_p\left(\hat{\psi}|m\right)\right] \end{aligned}$$

or we can directly expand the unconditional estimation in the definition of the variance

$$\begin{aligned} V\left(\hat{\psi}\right) &= E\left[\left(\hat{\psi} - E\left[\hat{\psi}\right]\right)^2\right] = E_p\left[E_m\left[\left(\hat{\psi} - E_p\left[E_m\left[\hat{\psi}|s\right]\right]\right)^2|s\right]\right] \\ &= E_m\left[E_p\left[\left(\hat{\psi} - E_m\left[E_p\left[\hat{\psi}|m\right]\right]\right)^2|m\right]\right] \end{aligned}$$

This form of direct expansion is particularly useful when combined with variance decomposition techniques.

An alternative to design-based inference in sample surveys is model based inference which assumes the actual values, both observed and unobserved, for each element in the finite population of interest are just one possible realization out of a conceptual superpopulation. In this case each element in the finite population can be modelled as a random variable and that the set spanned by the range of all of these random variables is called the *superpopulation*. Lohr (1999, p 47) points out that the joint probability distribution of these random variables

'*...supplies the link between units in the sample and units not in the sample in this model-based approach - a link that is missing in the randomization approach.*'

In the model based approach expectations are taken over repeated realizations of the finite population, $E_M[.]$. Hence our estimator, $\hat{\psi}$ would be model unbiased if

$$E_M\left[\hat{\psi} - \psi\right] = 0$$

and the estimator is now also a function of the random variables comprising the superpopulation from which the finite population of interest is one possible realization. The Mean Square Error of this estimator then becomes

$$E_M\left[(\hat{\psi}-\psi)^2\right] = 0$$

For more detail on the model-based approach to survey inference see Brewer (1963), Valliant *et al.* (2000) and Royall and Herson (1973). There remain some philosophical arguments between those who favour the model and design based approaches to inference in sample surveys. Sarndal *et al.* (1992, p 537) highlight the viewpoint for those who favour the model-based approach,

'*The model-based approach offers an alternative to design-based inference for survey populations. Advocates of the model-based approach do not reject randomization selection per se. However, they do oppose the use of the distribution on the grounds that such inferences, although distribution free, refer to repeated draws of samples, instead of to the particular sample that was actually drawn.*'

In comparison proponents of the design-based approach feel that the hypothetical superpopulation postulated by users of model-based inference, is an unrealistic and unnecessary construct. Smith (1994, p 8) suggests that the model-assisted approach has lead to some reconciliation between supporters of these two approaches,

'*A model-assisted estimator is chosen to satisfy both model-based and p-based criteria. For example, a regression estimator may be chosen because of the anticipated population structure, and it can then be made approximately p-unbiased.*'

There has been growing support for the model-assisted approach with proponents such as Brewer (1979); Hansen *et al.* (1983) and Sarndal *et al.* (1992).

2.1.2 Variance Decomposition Techniques

The Mean Square Error (MSE) or its component bias and variance should be provided to users to summarize the level of error that can be associated with statistics estimated from sample surveys. MSE decomposition techniques can then be applied

to split the variance of a statistic into a number of further components that can be attributed to various sources. Although Platek and Sarndal (2001) suggest that variance decomposition techniques become overly complex and unworkable when applied to practical situations, these techniques are still very powerful when examining a specific component of the total survey error, such as the interviewer effect.

Notation

This section will specify the notation that can be used to examine the interviewer effect. Concepts such as the *true* value will not be considered in detail here, but will be defined when applied in a particular model.

Consider a sample, s, of n elements drawn from a finite population, u, containing N units. Assume each element i, in the sample, s, is interviewed only once and in that one case by interviewer j for each trial, or realization of the survey, t. Then let

- y_{ijt} be the observed value for a characteristic of interest measured on element i by interviewer j during survey trial t.
- Y_i be the *true* value for the characteristic of interest for element i. The *true* value is independent of the sampling design, interviewer and particular trial in which the sample was observed.
- $\bar{Y}_s = \frac{1}{n}\sum_{i \in s} Y_i$ be the *true* sample mean of the characteristic of interest. The subscript s is used to indicate that the statistic has been aggregated over all elements selected in the sample set s.
- $\bar{Y}_u = \frac{1}{N}\sum_{i \in u} Y_i$ be the *true* population mean of the characteristic of interest. The subscript u is used to indicate that the statistic has been aggregated over all elements selected in the frame population set u.
- $\bar{y}_{s,t} = \frac{1}{n}\sum_{i \in s} y_{ijt}$ be the sample mean of the characteristic of interest for a given trial or realization, t, of the sample survey.

Then the total Mean Squared Error (MSE) of the sample mean $\bar{y}_{s,t}$, as an estimate of $\bar{Y}_u$ can be decomposed using variance decomposition techniques introduced

by Hansen *et al.* (1953). The following decomposition is an adaption of work presented by Lessler and Kalsbeek (1992, p 283)

$$\begin{aligned}
MSE\left(\bar{y}_{s,t}\right) &= E\left[\left(\bar{y}_{s,t}-\bar{Y}_u\right)^2\right]+B^2 \\
&= E\left[\left(\bar{y}_{s,t}-\bar{Y}_s+\bar{Y}_s-\bar{Y}_u\right)^2\right]+B^2 \\
&= E\left[\left(\bar{y}_{s,t}-\bar{Y}_s\right)^2\right]+E\left[\left(\bar{Y}_s-\bar{Y}_u\right)^2\right]+2E\left[\left(\bar{y}_{s,t}-\bar{Y}_s\right)\left(\bar{Y}_s-\bar{Y}_u\right)\right]+B^2 \\
&= MV+SV+CSM+B^2
\end{aligned}$$

where these terms attempt to deal separately with the various random processes that contribute to the survey error. They can be interpreted as, respectively,

- the measurement variance (MV), dealing with the stochastic nature of the measurement process. The interviewer effect can generally be viewed as a further sub-component of the measurement variation term. This component would disappear if we observed the *true* values, or if there was no non-sampling error

- the sampling variance (SV), dealing with the uncertainty introduced by the stochastic nature of the sampling process. This component will disappear if the entire finite population is observed, as in a census

- the covariance between sampling and measurement variance (CSM), caused by any correlation between factors influencing the sampling design and covariates in the measurement error model. This term will disappear if the sample design is ignorable

- a squared bias term (B)

Note that if the *true* population mean for the characteristic of interest $\bar{Y}_u$ is defined subjectively this will only affect the bias term, B. This is because any difference between our operational *true* value $\bar{Y}_u$ and another possible definition of the *true* value (which we will refer to as $\bar{\mathcal{Y}}_u$) will fall out of the decomposition as follows

$$\begin{aligned}
MSE\left(\bar{y}_{s,t}\right) &= E\left[\left(\bar{y}_{s,t}-\bar{\mathcal{Y}}_{u}\right)^{2}\right] \\
&= E\left[\left(\bar{y}_{s,t}-\bar{Y}_{u}+\bar{Y}_{u}-\bar{\mathcal{Y}}_{u}\right)^{2}\right] \\
&= E\left[\left(\bar{y}_{s,t}-\bar{Y}_{u}\right)^{2}\right]+E\left[\left(\bar{Y}_{u}-\bar{\mathcal{Y}}_{u}\right)^{2}\right]+2E\left[\left(\bar{y}_{s,t}-\bar{Y}_{u}\right)\left(\bar{Y}_{u}-\bar{\mathcal{Y}}_{u}\right)\right] \\
&= E\left[\left(\bar{y}_{s,t}-\bar{Y}_{u}\right)^{2}\right]+E\left[\left(B^{*}\right)^{2}\right]+2B^{*}E\left[\left(\bar{y}_{s,t}-\bar{Y}_{u}\right)\right] \\
&= V\left(\bar{y}_{s,t}\right)+\left(B^{*}\right)^{2}
\end{aligned}$$

where the bias term, B^* above is the difference between an alternative operationally defined *true* value and the original, objective *true* value for the characteristic of interest. Hence one advantage of using the MSE variance decomposition for the estimation of the measurement variance is that the *true* value of the characteristic of interest only needs to be known when estimating the bias as part of total survey error and not in the estimation of the measurement variance. This was recognized by Hansen *et al.* (1961, p 362) who stated

'*The unknown true value does not affect the response or sampling variance, but only the bias*'

One possible definition of Y_i is the expectation of the observed value of a characteristic of interest y for an element i over all possible trials (and therefore also under all possible sample designs, denoted here by $E_i[.|i]$). Given element i we can define an objective *true* value of response Y_i as $E_i[y_{ijt}|i]$. In which case any discrepancy between any other operationally defined *true* value $\mathcal{Y}_i$ and the unknown objective *true* value, Y_i, will be captured by the bias term B in the variance decomposition.

It is possible to construct even more elaborate and detailed decomposition structures than just the three terms in the decomposition presented above. Koch (1973), for example, splits the measurement variance term into three further categories; the Simple Response Variance, the Simple Correlated Response Variance and the Interaction Response Variance. However, the more branches that are added to this decomposition tree, the harder it generally becomes to relate these terms to real-world effects. These variance decomposition models applied in the literature do not generally adopt standard terminology and consequently it can be difficult to

compare decomposition terms across studies. Some examples of the discrepancies between variance decomposition studies can be found in Lessler and Kalsbeek (1992, Chapter 11).

Using variance decomposition techniques to formulate a total survey error model requires complete specification of all possible sources of error. Consequently, although variance decomposition methods have greatly influenced research into total survey error models their primary application has been the examination of specific sources of error. Variance decomposition models including a variety of formulations of the US Bureau of the Census Model, combined with Mahalanobis' technique of interpenetration (Mahalanobis, 1946), dominated research into the interviewer effect prior to the mid 1980s.

2.1.3 Correlational Approach

A number of different variance decomposition models have been specified for the examination of measurement error. In the US Bureau of the Census model the characteristics and actions of the interviewer are assumed to introduce a positive correlation between the responses of elements collected by each interviewer. Thus the effect of the interviewer is to introduce dependence among observations in the same workload resulting in an increase in the variance of estimates. This will lead to inappropriate inference if the interviewer effect on survey statistics is not incorporated into the variance estimation procedure.

O'Muircheartaigh (1997) refers to this model as the correlational approach. It is adopted by authors such as Koch (1973) and Koop (1974). The Bureau of the Census model is a term which has been applied by authors such as Bailar (1976); Lessler and Kalsbeek (1992) to refer to the class of variance decomposition response error models that were developed in a number of articles by Hansen et al at the US Bureau of the Census (Hansen and Hurwitz, 1943; Hansen *et al.*, 1953, 1961). These articles developed a general model for characterizing the various components of response error.

Notation and Assumptions

This section will introduce some further notation and some of the assumptions that have been applied under the correlational approach for examining the interviewer effect.

Assume that a sample of n population elements are selected out of N possible elements. These elements are allocated to one of k interviewers such that each element i is interviewed only once by interviewer j. Let g_j indicate the set of all n_j elements that interviewer j enumerates. Hence the sum over all interviewers of the interviewer allocations will contain the same elements without repetition as the sample, $\bigcup_{j=1}^{k} g_j = \{g_1, \ldots, g_k\} = \{1, \ldots, n\} = s$. A summation over all the elements in the sample can be described as $\sum_{i \in s} y_i = \sum_{j=1}^{k} \sum_{i \in g_j} y_i$. It is assumed that observations collected by the same interviewer will have correlation ρ, while observations collected by different interviewers are not correlated.

Under the correlational approach it is also common to make explicit assumptions regarding the sample/interviewer selection and allocation scheme. For example Hansen *et al.* (1951, p 157) outline their model as follows

(a) *'n of the N individuals in the population are selected at random without restriction.*

(b) *k_A interviewers are selected at random without restriction from the A-th interviewer group to interview those sample individuals who are available for interview by this interviewer group. Let $k = \sum_{A}^{L} k_A$ be the total number of interviewers selected.*

(c) *The same number, $\bar{n}$, of individuals is assigned to each of the k_A interviewers. The $\bar{n}$ individuals assigned to any interviewer are a random subsample of all*

the sample individuals available for interview by this interviewer group.'

In the following derivation it has been assumed that each interviewer j has been assigned an allocation of respondents g_j of size n_j and y_{ijt} is the observed value of a characteristic of interest, y, on individual i, by interviewer j, in trial t. Then assume the *true* value of the characteristic of interest under the correlational model is $E_m[y_{ijt}] = Y_i$, the variance of the observed values is $V_m[y_{ijt}] = \sigma^2$ and different observations are only correlated if they are collected by the same interviewer. I.e.

$$Cor_m[y_{ijt}.y_{i'j't}|s] = \begin{cases} 1 & \text{if } i = i', j = j' \\ \rho & \text{if } i \neq i', j = j' \\ 0 & \text{otherwise} \end{cases}$$

Define the individual measurement deviation d_{ijt} for an element i interviewed by j in a given trial, t, to be the difference between the observed and the *true* value for a characteristic of interest, i.e. $d_{ijt} = y_{ijt} - Y_i$, then

$$\begin{aligned} E_m[d_{ijt}|s] &= E_m[y_{ijt} - Y_i|s] = 0 \\ E_m[d_{ijt}.d_{i'jt}|s] &= Cov_m[d_{ijt}.d_{i'jt}|s] + E_m[d_{ijt}|s]\, E_m[d_{i'jt}|s] \\ &= \sigma^2\rho \\ V_m[d_{ijt}] &= V_m[y_{ijt}|s] \\ &= \sigma^2 \end{aligned}$$

Then applying these results the measurement variance can be shown to be

$$
\begin{aligned}
MV\left(\bar{y}_{s,t}\right) &= E\left[\left(\bar{y}_{s,t}-\bar{Y}_s\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum\nolimits_{j=1}^{k}\sum\nolimits_{i\in g_j} d_{ijt}\right)^2\right] \\
&= \frac{1}{n^2}E\left[\sum\nolimits_{j=1}^{k}\sum\nolimits_{i\in g_j} d_{ijt}^2+\sum\nolimits_{j=1}^{k}\sum\nolimits_{i\neq i'\in g_j} d_{ijt}.d_{i'jt}\right] \\
&+ \frac{1}{n^2}E\left[\sum\nolimits_{j\neq j'}^{k}\sum\nolimits_{i\in g_j}\sum\nolimits_{i'\in g_{j'}} d_{ijt}.d_{i'j't}\right] \\
&= \frac{1}{n^2}E_p\left[\sum\nolimits_{j=1}^{k}\sum\nolimits_{i\in g_j} E_m\left[d_{ij}^2|s\right]+\sum\nolimits_{j=1}^{k}\sum\nolimits_{i\neq i'\in g_j} E_m\left[d_{ij}.d_{i'j}|s\right]\right] \\
&= \frac{\sigma^2}{n^2}E_p\left[\sum\nolimits_{j=1}^{k} n_j+\rho\sum\nolimits_{j=1}^{k} n_j\left(n_j-1\right)\right] \\
&= \frac{\sigma^2}{n}+\frac{\sigma^2\rho}{n^2}E_p\left[\sum\nolimits_{j=1}^{k} n_j^2-\sum\nolimits_{j=1}^{k} n_j\right] \\
&= \frac{\sigma^2}{n}-\frac{\sigma^2\rho}{n}+\frac{\sigma^2\rho}{n^2}E_p\left[k\sigma_{n_j}^2+\sum\nolimits_{j=1}^{k}\left(\frac{n}{k}\right)^2\right] \\
&= \frac{\sigma^2}{n}\left[1+\left\{\frac{n}{k}\left(1+E_p\left[CV_{n_j}^2\right]\right)-1\right\}\rho\right] \qquad (2.3)
\end{aligned}
$$

Where $\bar{n} = \frac{n}{k}$, $CV_{n_j}^2 = \frac{\sigma_{n_j}^2}{\bar{n}^2}$ is the relative variance of the size of the interviewer allocations, n_j, and $\sigma_{n_j}^2$ is the variance of n_j. Then for a given sample size, n, level of correlation, ρ, and measurement variation, σ^2, MV will be minimized when all interviewers are allocated an equal number of respondents, $\bar{n}$. In this case this will reduce to $\frac{\sigma^2}{n}\left[1+\left(\bar{n}-1\right)\rho\right]$, the standard result derived by Hansen *et al.* (1951).

Summary

This general correlational approach can be extended or adapted to explore specific topics of interest. For example Felligi (1964, 1974) improved estimation techniques, by considering the case of repeated surveys, for the correlated response variance while Koch (1973) extended the US Bureau of the Census' response error model to the multivariate case, while Lessler and Kalsbeek (1992) applied variance decomposition methods to the formulation of a total survey error model. The strength of this approach is its generality, as it can be applied either as a total survey error model

or to explore a specific error component such as the interviewer effect.

In practice it has proven difficult to estimate the interviewer effect using these models because the groups of respondents allocated to interviewers tend to be geographically concentrated. If this is the case the interviewer effect will be confounded by the correlation introduced by the geographic proximity of observations. Consequently it may be too costly to produce estimates of the interviewer effect as some form of replication of measurements in each spatial region would be required for the separation of the confounded effects. Mahalanobis' technique of interpenetration (Mahalanobis, 1946) is still considered the minimum requirement for the estimation of the interviewer effect. However, if it is assumed that there is no memory build-up, repeated measurement of spatial regions can also be applied as a solution to this problem (Felligi, 1964).

2.1.4 The ANOVA Model Approach

Another approach to examining the interviewer effect is to assume that the observed information collected from a single survey element i can be viewed as a combination of fixed and random effects. This approach has been developed by Kish (1962); Sukhatme and Sukhatme (1970); Hartley and Rao (1978) and has also been referred to as the '*Linear Models Approach*' by Lessler and Kalsbeek (1992) as it is based on linear models as described by Searle (1971).

Notation and Assumptions

Extending our earlier notation, let ϕ_{jt} be a random effect influencing the observed response y_{ijt} of any population element i interviewed by interviewer j due to the characteristics and actions of the interviewer during trial t, such that $\phi_{jt} \sim (0, \sigma^2_\phi)$ and $E[\phi_{jt}.\phi_{j't}] = 0$ for $j \neq j'$. Let ε_{ijt} be a random error term for the remaining discrepancy between the observed and *true* values for a characteristic of interest y such that $\varepsilon_{ijt} \sim (0, \sigma^2_\varepsilon)$ and $E[\varepsilon_{ijt}.\varepsilon_{i'j't}] = 0$ for $i \neq i'$.

Basic Derivations

Consider the following basic ANOVA model to describe our observed response, y,

$$y_{ijt} = Y_i + \phi_{jt} + \varepsilon_{ijt} \tag{2.4}$$

then

$$var(y_{ijt}) = var(Y_i) + \sigma^2_{int} + \sigma^2_{\varepsilon}$$

and if the expectation was first taken over the model, given a sample taken in trial t, $V_m(Y_i|s)$ would be zero and subsequently the total variance would be $V_m(y_{ijt}|s) = \sigma^2 = \sigma^2_{\phi} + \sigma^2_{\varepsilon}$. Then it follows that two observations made on distinct individuals i and i' in the same g_j have covariance

$$\begin{aligned} Cov\left(y_{ijt}, y_{i'jt}\right) &= E\left[y_{ijt}.y_{i'jt}\right] - E\left[y_{ijt}\right]E\left[y_{i'jt}\right] \\ &= E\left[Y_i.Y_{i'} + \phi^2_{jt}\right] - E\left[Y_i\right]E\left[Y_{i'}\right] \\ &= Cov\left(Y_i, Y_{i'}\right) + \sigma^2_{int} \end{aligned}$$

Under our model, and conditional on the selected sample, s, there will be no covariance between the *true* values of the characteristic of interest y and hence observations collected by the same interviewer will have a covariance of σ^2_{int}. Therefore the intraclass correlation, ρ, between two elements allocated to the same interviewer can be expressed as

$$\begin{aligned} \rho &= \frac{Cov_m\left(y_{ij}, y_{i'j}|s\right)}{\sqrt{V_m\left(y_{ij}|s\right).V_m\left(y_{i'j}|s\right)}} \\ &= \frac{\sigma^2_{int}}{\sigma^2_{int} + \sigma^2_{\varepsilon}} \end{aligned}$$

we can then summarize the correlation between two observations y_{ijt} and $y_{i'j't}$, for a given trial t as follows

$$Cor_m\left(y_{ij}y_{i'j'}|s\right) = \begin{cases} \rho & \text{if } j = j' \text{and } i \neq i' \\ 0 & \text{otherwise} \end{cases}$$

Kish (1962) refers to ρ as '*the coefficient of intraclass correlation*'.

These are the same assumptions applied in the correlational approach in Section 2.1.3 and hence we can again show under the ANOVA approach that the Measurement Variance (MV) of the sample mean will be $\frac{\sigma^2}{n}\left[1+\left\{\frac{n}{k}\left(1+E_p\left[CV^2_{n_j}\right]\right)-1\right\}\rho\right]$.

Summary

The ANOVA models approach to examining the interviewer effect has proven less prevalent than the correlational models approach in the literature. O'Muircheartaigh (1997, p 8) recognizes one possible reason for this

'*...one drawback is that ANOVA models do not easily lend themselves to the analysis of categorical data.*'

The correlational approach is a very general methodology for examining survey errorThe ANOVA approach lends itself to the examination of specific sources of survey error which the user believes can be represented as directly interpretable random effects. Both models have similar aims and have been applied with various assumptions and definitions by a number of different authors. For summaries see Lessler and Kalsbeek (1992); O'Muircheartaigh (1997).

Both the ANOVA and the correlational approaches have been used in practice for estimating the interviewer effect in interpenetrated surveys. These techniques have since been generalized under a single multilevel or Generalized Linear Mixed Modelling (GLMM) approach. Up to this point we have only considered the impact of the interviewer effect on the measurement variance component of the total variance. The following section will consider how the interviewer effect impacts on the total variance and its estimate. We will then be able to consider the relevance of the interviewer effect in comparison with other sources of survey error.

2.2 Estimating the Total Variance

In the previous section we derived the measurement error component of the total variance of the sample mean based on a simple interviewer effect model. Result (2.3) provides a clear indication of the impact of the interviewer effect on the Measurement Variance (MV), but does not consider the impact of the interviewer on the Total Variance (TV). Based on our variance decomposition model it is however straightforward to determine the impact of the interviewer effect on the Total Variance. For example if we assume the sampling design is ignorable for the measurement error, the covariance between the sampling and measurement variance (CSM) will be zero and the total variance of the sample mean will be $\operatorname{var}(\bar{y}_{s,t}) = MV(\bar{y}_{s,t}) + SV(\bar{y}_{s,t})$. Then for equal allocations of respondents to interviewers we can say under Model (2.4) and under Simple Random Sampling With Replacement (SRSWR) the total variance of the sample mean will be

$$TV(\bar{y}_{s,t}) = \frac{S_Y^2}{n} + \frac{(\sigma_\varepsilon^2 + \sigma_{int}^2)}{n}\left[1 + (\bar{n} - 1)\rho\right] \tag{2.5}$$

where $S_Y^2 = \frac{1}{N-1}\sum_{i\in u}\left(Y_i - \bar{Y}_u\right)^2$ is the adjusted population variance of the *true* population values and ρ is the intra-interviewer correlation coefficient, $\rho = \frac{\sigma_{int}^2}{\sigma_\varepsilon^2 + \sigma_{int}^2}$.

In practice we do not observe the entire population and a common estimate of the variance of the sample mean under SRSWR which ignores any measurement error would be

$$\hat{\operatorname{var}}(\bar{y}_s) = \frac{s_y^2}{n} \tag{2.6}$$

where s_y^2 is the adjusted variance of the observed sample $s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)$. If there is no measurement error then $y_i = Y_i$ and our estimate (2.6) will be unbiased as $E\left[\hat{\operatorname{var}}(\bar{y}_{s,t})\right] = \frac{S_Y^2}{n} = \operatorname{var}(\bar{y}_{s,t})$. Similar results hold for other sampling schemes, though the form of the estimate will change, for example under Simple Random Sampling Without Replacement (SRSWOR) the estimator $\hat{\operatorname{var}}(\bar{y}_{s,t}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}$ will be unbiased if there is no measurement error.

When we are faced with measurement error, for example in Model (2.4), esti-

mates which ignore the measurement error such as (2.6) are no longer unbiased. We can demonstrate this by deriving the expectation of the adjusted variance of the observed sample under SRSWR for measurement error model (2.4).

$$
\begin{aligned}
E\left[s_y^2\right] &= E\left[\frac{1}{n-1}\sum\nolimits_{i\in s}(y_i-\bar{y}_s)^2\right] \\
&= \frac{1}{n}E\left[\sum\nolimits_i y_i^2\right] - \frac{1}{n\,(n-1)}E\left[\sum\nolimits_{i\neq i'} y_i y_{i'}\right] \\
&= \frac{1}{n}E_p\left[E_m\left[\sum\nolimits_i (Y_i+\phi_j+\varepsilon_{ij})^2|s\right]\right] \\
&- \frac{1}{n\,(n-1)}E_p\left[E_m\left[\sum\nolimits_{i\neq i'}(Y_i+\phi_j+\varepsilon_{ij})\,(Y_{i'}+\phi_{j'}+\varepsilon_{i'j'})|s\right]\right] \\
&= \frac{1}{n}\left\{E_p\left[\sum\nolimits_i Y_i^2\right]+E_p\left[E_m\left[\sum\nolimits_i \phi_j^2|s\right]\right]+E_p\left[E_m\left[\sum\nolimits_i \varepsilon_{ij}^2|s\right]\right]\right\} \\
&- \frac{1}{n\,(n-1)}\left\{E_p\left[\sum\nolimits_{i\neq i'} Y_iY_{i'}\right]+E_p\left[E_m\left[\sum\nolimits_j n_j\,(n_j-1)\,\phi_j^2|s\right]\right]\right\} \\
&= \frac{1}{n}\sum\nolimits_{i\in u} Y_i^2 E_p\left[U_i^2\right]+\sigma_\varepsilon^2+\sigma_{int}^2\left[1-\frac{n(CV_{n_j}^2+1)-k}{k(n-1)}\right] \\
&- \frac{1}{n\,(n-1)}\sum\nolimits_{i\neq i'\in u} Y_iY_{i'}E_p\left[U_iU_{i'}\right] \\
&= \frac{1}{N}\sum\nolimits_{i\in u} Y_i^2-\frac{1}{N\,(N-1)}\sum\nolimits_{i\neq i'\in u} Y_iY_{i'}+\sigma_\varepsilon^2+\sigma_{int}^2\left[\frac{n(k-1-CV_{n_j}^2)}{k(n-1)}\right] \\
&= S_Y^2+\sigma_\varepsilon^2+\sigma_{int}^2\left[\frac{n(k-1-CV_{n_j}^2)}{k(n-1)}\right]
\end{aligned}
$$

Under equal allocations of $\bar{n}=\frac{n}{k}$ respondents to each interviewer, j, this will be equal to

$$
E\left[s_y^2\right]=S_Y^2+\sigma_\varepsilon^2+\sigma_{int}^2\left[\frac{n-\bar{n}}{n-1}\right] \tag{2.7}
$$

We can then see that faced with an individual and interviewer level measurement error, as specified in Model (2.4) and with a SRSWR sampling scheme, estimates such as (2.6) will under-estimate the total variance of the sample mean (2.5) as $E\left[\hat{\text{var}}(\bar{y}_{s,t})\right]-\text{TV}(\bar{y}_{s,t}) = \sigma_{int}^2\left[\frac{k-n}{k(n-1)}\right]$ and $n>k>1$. We see that (2.6) appropriately reflects the contribution of the sampling error and the individual level measurement error to the total variance but underestimates the impact of the interviewer error on the total variance of the sample mean.

We must therefore adjust our estimate of the variance of the sample mean for the impact of the interviewer effect. This can be done by utilizing variance decomposition models to estimate the interviewer level intra-class correlation coefficient based on the total variation observed in the sample. Practical application of variance decomposition models to estimate the intra-interviewer correlation will be demonstrated in Chapter 3 while the theory underlying multilevel variance decomposition models will be discussed in the remainder of this chapter. However the total variance in the observed sample will be $\mathrm{var}(y_i) = S_Y^2 + \sigma_{int}^2 + \sigma_\varepsilon^2$ and hence in practice we can only estimate $\rho^* = \frac{\sigma_{int}^2}{\sigma_{int}^2+\sigma_\varepsilon^2+S_Y^2}$ rather than $\rho = \frac{\sigma_{int}^2}{\sigma_{int}^2+\sigma_\varepsilon^2}$ with these models. Note that we can then rearrange (2.7) to express the adjusted sample variance in terms of $\sigma_C^2 = \sigma_\varepsilon^2 + S_Y^2$ and ρ^*.

$$\begin{aligned} E[s_y^2] &= \sigma_C^2 + \sigma_{int}^2 \left[\frac{n-\bar{n}}{n-1}\right] \\ &= \sigma_C^2 + \sigma_C^2 \left[\frac{\rho^*}{1-\rho^*}\right]\left[\frac{n-\bar{n}}{n-1}\right] \\ &= \sigma_C^2 \left[\frac{n-1+\rho^*-\bar{n}\rho^*}{(1-\rho^*)(n-1)}\right] \end{aligned}$$

This holds as $\sigma_{int}^2 = \sigma_C^2 \frac{\rho^*}{1-\rho^*}$. Then given an estimate $\hat{\rho}^*$ of ρ^* we can produce the following estimate of σ_C^2

$$\hat{\sigma}_C^2 = s_y^2 \left[\frac{(1-\hat{\rho}^*)(n-1)}{n-1+\hat{\rho}^*-\bar{n}\hat{\rho}^*}\right] \tag{2.8}$$

and then we can also rearrange (2.5) to express the total variance of the sample mean in terms of ρ^* and σ_C^2.

$$\begin{aligned} \mathrm{var}\left(\bar{y}_{s,t}\right) &= \frac{S_Y^2}{n} + \frac{(\sigma_\varepsilon^2+\sigma_{int}^2)}{n}\left[1+(\bar{n}-1)\rho\right] \\ &= \frac{1}{n}\left\{\sigma_C^2+\sigma_{int}^2+(\bar{n}-1)\sigma_{int}^2\right\} \\ &= \frac{1}{n}\left\{\sigma_C^2+\bar{n}\sigma_C^2\frac{\rho^*}{1-\rho^*}\right\} \\ &= \frac{\sigma_C^2}{n}\left\{1+\bar{n}\frac{\rho^*}{1-\rho^*}\right\} \end{aligned}$$

and then given an estimate $\hat{\rho^*}$ of ρ^* we can produce the following estimate of the total variance of the sample mean, $\bar{y}_{s,t}$

$$\begin{aligned} \hat{\text{var}}\left(\bar{y}_{s,t}\right) &= \frac{\hat{\sigma}_C^2}{n}\left\{1+\bar{n}\frac{\hat{\rho}^*}{1-\hat{\rho}^*}\right\} \\ &= \frac{s_y^2}{n}\left[\frac{(n-1)\{1+(\bar{n}-1)\hat{\rho}^*\}}{n-1-(\bar{n}-1)\hat{\rho}^*}\right] \end{aligned} \tag{2.9}$$

For unequal allocations of respondents to interviewers this result becomes

$$\text{v}\hat{a}r(\bar{y}_{s,t}) = \frac{s_y^2}{n}\left[\frac{(n-1)\{1+(\frac{n}{k}\{1+CV_{n_j}^2\}-1)\hat{\rho}^*\}}{n-1-(\frac{n}{k}\{1+CV_{n_j}^2\}-1)\hat{\rho}^*}\right] \tag{2.10}$$

Previous studies examining the relative impact of the interviewer effect and the survey design on estimates derived from surveys, such as O'Muircheartaigh and Campanelli (1998), have considered the interviewer effect as being synonymous with the design effect (DEFF) and apply $1+(\bar{n}-1)\rho^*$ as an estimate of the inflationary impact of the interviewer effect. This is an intuitively appealing estimate as we can re-arrange (2.5) to show that

$$TV\left(\bar{y}_{s,t}\right) = \frac{\sigma_T^2}{n}\left[1+\left(\bar{n}-1\right)\rho^*\right] \tag{2.11}$$

where $\sigma_T^2 = \sigma_\varepsilon^2 + S_Y^2 + \sigma_{int}^2$. However for alternative non-SRSWR sampling schemes the sampling variance component of the total variance of the mean will not be $\frac{S_Y^2}{n}$ and we will not be able to simplify the total variance to produce (2.11). Furthermore applying (2.11) to estimate the inflationary impact of the interviewer effect is inappropriate in practice as it does not consider the impact of the interviewer effect on estimates of the total variance. We can see from (2.9) that for SRSWR, use of (2.11) would result in an underestimate of the inflationary impact of the interviewer effect on our *estimate* of the total variance of the sample mean by a factor of $\frac{n-1}{n-1-(\bar{n}-1)\rho^*}$.

We have derived a new estimate of the total variance of the sample mean that explicitly considers the impact of sampling and measurement as distinct processes. We can also see that (2.9) will remain an unbiased estimate of total variance of the sample mean when there is no individual level measurement error. However this

expression will only approximate the variance of the sample mean for a SRSWOR sampling scheme. Assuming there is no individual level measurement error, similar results can be derived to consider the impact of the interviewer effect under any alternate sampling scheme. In contrast when there is both an interviewer and an individual level measurement error, we can only estimate the total variance under SRSWR unless we have further information regarding the magnitude of the individual level measurement variance, σ_ε^2. This occurs because the measurement variance only relates to the sample and is in general (for example if the sample design is ignorable) not influenced by the sampling design.

Given that s_y^2 can be calculated from the sample we only need an estimate of ρ^* to calculate either (2.9) or (2.10) and therefore produce an estimate of the total variance of the sample mean. We have seen how ANOVA and correlational approaches can be used to provide an estimate of ρ^*. These approaches have, however, proven limiting for large-scale surveys. The following section will discuss a more general GLMM variance decomposition model that can be used to estimate ρ^* and show how design information can be incorporated under this approach.

2.3 Multilevel Modelling Approach to Estimating the Interviewer Effect

Both the simple response model and the simple sampling design scheme discussed earlier can be generalized to consider more complex scenarios. Lessler and Kalsbeek (1992) show that the indicator variable (2.1) only considers the influence of the sampling design and introduce further indicator variables which cater for other, non-sampling, aspects of the survey design. When considering the interviewer effect Lessler and Kalsbeek (1992) define two further indicator variables, one to indicate the selection of an interviewer out of a finite population of interviewers and another to consider the allocation scheme of respondents to interviewers, in order to incorporate the stochastic nature of these processes into their survey error model.

As an extension to the simple response error model (2.2) and the ANOVA response model approach (2.4), Hox (1994) showed that multilevel response models can be considered to be the most appropriate for considering data collected from sample surveys in which respondents are nested within higher level groupings.

We will now introduce the multilevel response model and how it can be used to consider the interviewer effect. The multilevel model is a succinct approach to describing the dependency structure within the data and facilitates estimation of variance components such as the interviewer effect. Early multilevel studies followed an ANOVA style representation of the data in which the indexation of elements was nested. This is equivalent to the notation presented in (2.3). This indexation can be misleading however for complex multilevel structures in which the data is highly non-hierarchical. Browne *et al.* (2001) and Rasbash and Browne (2001) present an alternative notation for multilevel modelling which clarifies the nesting structure of the data. However an ANOVA style nested indexation notation can still be applied to succinctly summarize the data provided the classification structure of the data is also explicitly demonstrated. For this purpose the classification diagram of Browne *et al.* (2001) is a flexible tool that will be applied in this thesis. In this way a succinct, suitable representation of a multilevel system can be provided.

The multilevel ANOVA style representation of multilevel systems does not explicitly describe the link between estimation of the interviewer effect and the survey design. For these purposes the more general GLMM notation can be considered. The multilevel model is a subset of the GLMM that presents a more concise summary in practice. In this thesis multilevel and GLMM notations will be presented interchangeably.

In the following we introduce the multilevel and GLMM response models that will be applied to estimate the interviewer effect. Data structures that can be used to consider the interviewer effect will be described by way of example and further extensions to the response model will be considered. For the following section all expectations will be taken with respect to the relevant model and will be conditional on the selected sample.

2.3.1 Multilevel Modelling

In the real world human populations do not generally exhibit simple structures that lend themselves to the application of equally simplistic models. This was recognized by Jones (1992, p 238) who summarizes a theme of Skinner *et al.* (1989)

'*...that failing to take complex structure into account when estimating statistical models leads to severe technical problems.*'

The complex nature of many human populations may often be simplified by specifying the hierarchical structure of the individual elements in the population. This structure may be a simple linear hierarchy, for example, people can be grouped into households, which can further be grouped into suburbs while the set of all suburbs may be considered the population of a town. We can also conceive of multiple membership or cross-classified hierarchical structures in which a single element may belong to several different hierarchies at the same level. For example a population of teenagers in a single town could be classified according to either their classes and schools or according to the households and suburbs in which they live. As an extension to this example we can also conceive of the multiple membership situation in which a single teenager, with divorced parents, belongs to two households simultaneously.

Elements in the same hierarchical grouping will generally be more similar than elements in different groups. This can be viewed as a two-way process since individuals with similar characteristics will tend to cluster together while individuals in the same group will also be exposed to one another and become more similar over time. Thus the complex structure of a population leads to a dependency structure that may invalidate independence assumptions commonly applied in fixed effect models. Since the mid-1980s, multilevel modelling techniques have been developed to take the complex structure of the population into account and produce appropriate parameter estimates, see for example Goldstein (1995).

Multilevel Response Models Applied to Estimating the Interviewer Effect

A good summary of the state of knowledge regarding the interviewer effect prior to the introduction of multilevel modelling techniques is provided by Collins (1980). He provides a critical literature review and suggests some partial solutions to the problem. In general the studies he describes ignore the complex structure of the population of interest. Collins (1980, p 83) identifies this as a weakness in the interviewer effect literature that can lead to further confounding of the interviewer effect with other effects in the population

'*Where the division into sub-classes leads to comparisons between different sets of interviewers, the interviewer effect is confounded with the sub-class comparison. The most obvious example is a regional analysis; a less obvious example is a longitudinal comparison, if measurements at different points in time are made by different interviewers. In such cases, an apparent area or time effect could actually represent only the interviewer effect.*'

Around the same time as Collins (1980) review, a number of authors were developing new models for the examination of the interviewer effect. They attempted to directly incorporate response effects into their models in order to more appropriately cater for the complex structure of the population in question. Aitken *et al.* (1981) applied variance component models, also known as hierarchical linear models, multilevel models or random component models, to examine the interviewer effect. O'Muircheartaigh and Wiggins (1981) also attempted to directly model the interviewer effect in a log-linear analysis of aircraft noise on respondent annoyance levels.

One of the first major studies applying multilevel modelling techniques to the estimation of the interviewer effect was undertaken by Anderson and Aitken (1985). They derived the maximum likelihood estimation procedure for variance component models with Bernouilli response and estimated the interviewer effect, including covariates, in a large survey carried out by Social and Community Planning and Research. In this study there were 1265 respondents with 2 interviewers allocated to each of the 32 areas in the study. As there were 2 interviewers allocated to each

area, the areas could be considered to be interpenetrated and hence random effects due to the area and the interviewer could be separately estimated. None of the variables indicated both significant interviewer and area effects.

Since the mid 1980s there have been consistent advances in the versatility and applicability of multilevel modelling techniques. A number of authors, such as Goldstein (1991) recognized that multilevel models can be applied to more appropriately consider the hierarchical structure of survey data. This notion was considered in further detail by Hox (1994) who showed that multilevel modelling techniques can be considered to be the most appropriate for dealing with data collected from survey designs in which respondents are nested within higher level groupings.

Further Extensions

Further theoretical extensions to multilevel models were provided by Raudenbush (1993); Rasbash and Goldstein (1994) who considered cross-classified and mixed hierarchical random structures. Raudenbush (1993) discussed the notion of applying crossed random effects model for longitudinal research and the theoretical aspects of this application were expanded in Goldstein (1995) and Snijders (1996); Yang and Goldstein (1996). Further discussion of how to handle discrete and dichotomous data through the application of multilevel logistic regression models, can be found in Goldstein *et al.* (1998) and Snijders and Bosker (1999).

With each successive advance adding to the versatility and applicability of multilevel modelling, parameter estimation has become more complicated and computer intensive. Snijders and Bosker (1999, p 219) describe some of the difficulties resulting from this increasing complexity stating

'*The estimation procedures for these models still are in a state of development. The choice between these procedures should be based on stability of the algorithm (will the algorithm converge to a valid estimate?), statistical efficiency, availability of software, and the possibility to obtain parameter tests.*'

Snijders and Bosker (1999) continue on to highlight first order marginal quasi-likelihood estimation, as implemented in the software packages MLwiN, HLM and

VARCL, to be the most stable and rapid of the currently available algorithms.

Increased Application of Multilevel Models to Estimate the Interviewer Effect

In the 1990's multilevel models became more popular for examining both survey data and the interviewer effect. Hox *et al.* (1991) discussed how studies of survey data that do not apply multilevel modelling techniques are more prone to type I error. However, despite applying multilevel modelling techniques Hox *et al.* (1991) only found marginal interviewer effects. In contrast, Pannekoek (1991) who directly analyzed the interviewer intraclass correlation using mixed models found evidence of a substantive interviewer effect in two out of five opinion questions. A number of studies have also applied multilevel models to analyze how the characteristics of the interviewer are related to the interviewer effect; see for example Ecob and Jamieson (1992); Beerten (1999); Martin and Beerten (2002).

An early summary of the applicability of multilevel models for survey analysis was presented by Jones (1992). He highlights several features that make multilevel models attractive, such as their ability to deal with autocorrelation and produce appropriate inference for survey data by taking into account its complex structure. However, despite these advantages, cost constraints associated with the running of surveys generally still result in the random effects due to the region being confounded with that of the interviewer. Consequently the aforementioned studies have assumed interpenetration or negligible regional effects, although Pannekoek (1991, p 525) does note

'*The respondents were not randomly assigned to interviewers. Therefore a random effect associated with interviewers cannot be interpreted as reflecting interviewer effects alone. Also regional differences in the response will contribute to the estimate of the random effect.*'

Despite this recognition Pannekoek (1991) assumes the 2 independent random subsamples have approximately the same contribution to regional differences and there-

fore that the estimate of the interviewer effect is appropriate.

As new multilevel modelling techniques have been developed, studies of the interviewer effect utilizing these models have also become more sophisticated. Paralleling the theoretical work of Goldstein (1991) and Hox (1994), Wiggins *et al.* (1992) considered a more explicit hierarchical structure incorporating covariates during a practical application of multilevel models for the analysis of the interviewer effect in two different studies. Both of these studies were conducted in concentrated geographic areas with random allocation of interviewers. The interpenetration in this case allowed the interviewer and the area-based random effects to be separately estimated and a substantial interviewer effect, constituting approximately 29% of the total variance was found.

Further extensions of the work of Wiggins *et al.* (1992); Rasbash and Goldstein (1994) to the analysis of the interviewer effect in cross-classified hierarchical structures were provided by O'Muircheartaigh and Campanelli (1998, 1999). They found significant interviewer effects in an interpenetrated sample experiment. Pickery (2000); Pickery and Loosveldt (2000) also extended cross-classified multilevel models to examine the interviewer effect in an longitudinal electoral survey conducted in a single region (Flemish) of Belgium. Pickery and Loosveldt (2001) later applied multilevel logistic regression techniques to consider dichotomous outcomes when examining the mediating influence of question characteristics on non-response, while Pickery and Loosveldt (2004) showed how multilevel modelling techniques can be applied to identify exceptional interviewers.

Summary

Recent studies applying multilevel modelling techniques to examine the interviewer effect have been applying more sophisticated and appropriate models to consider the complex nature of survey data. They have, however, yet to consider the situation in which the area-based effects are confounded with the interviewer effects due to cost constraints and geographically concentrated allocations of respondents to interviewers. Studies such as O'Muircheartaigh and Campanelli (1998, 1999) examined an

interpenetrated sample, while Pickery and Loosveldt (2000, 2001) consider a single concentrated geographic region in their study. The advantages of applying multilevel modelling techniques to survey data has been comprehensively demonstrated and the versatility of these models to a number of different situations demonstrated.

One of the weaknesses with the presentation of multilevel models is that the link between design and the estimability of the interviewer effect is implicit and hence not clear from the stated model. When considering interpenetrated designs it is therefore beneficial to adopt a more general notation that explicitly details the link between the design and the interviewer effect. The following section will demonstrate how all multilevel models can be presented under the broader GLMM framework and how the GLMM explicitly incorporates information regarding the design.

2.4 Generalized Linear Mixed Models (GLMMs)

The Generalized Linear Mixed Model (GLMM) is a broad class of models that cover all of the multilevel models introduced in this chapter. Whereas in the basic multilevel representation the design is implicitly incorporated within the stated model, the GLMM explicitly describes this information within the fixed and random effect design matrices. Hence, although there may be a single multilevel representation of a model for the estimation of the interviewer effect under a number of differing designs, there will be only one GLMM representation for each survey design and hierarchical classification structure combination.

When exploring the effect of interpenetrating designs on the estimation of the interviewer effect it is important to consider the design explicitly and this section will present a number of examples detailing how the interviewer effect can be considered under the GLMM. In general we will assume that the sample design is non-informative with respect to the measurement process and consider multi-stage survey designs. For simplicity of presentation ANOVA-style representations of the multilevel model will still be used, however these models will be conditional on an

implied survey design and presented alongside a classification diagram. Further information regarding GLMMs can be found in McCulloch and Searle (2001).

2.4.1 GLMMs: Background

Statistical techniques for considering data drawn from populations with hierarchical structures have been developed in a number of different disciplines under a variety of names. Nevertheless all of these techniques can be considered as specifications of the Generalized Mixed Model (GMM).

Mixed Models derive their name as they contain a *mixture* of both fixed and random effects. It then follows that random effects models are also a part of this family. In general the proportion of the total variance related to each random effect included in a generalized linear mixed model can be estimated as a variance component and the implications of variance components models have been discussed in detail by authors such as Searle *et al.* (1992).

Further subsets of the GMM have been defined through specification of the variance-covariance structure associated with the model. The hierarchical linear model of Byrk and Raudenbush (1992) is one such example where prior knowledge of the hierarchical structure of the data, and the dependency between observations that this implies, immediately leads to a block diagonal structure for the variance-covariance matrix. Similarly the multilevel modelling framework of Goldstein (1987), the random coefficient modelling framework of Longford (1993) and the Multiple Membership Multiple Classification (MMMC) framework of Browne *et al.* (2001) can also be developed through specification of the variance-covariance matrix in the GMM.

The next section will set out some of the assumptions of a simple generalized linear mixed model and derive by way of example the general structure of its variance-covariance matrix under a number of response structures appropriate for considering the interviewer effect.

2.4.2 The Generalized Linear Mixed Response Model

A Linear Mixed Model (LMM) for the response observed in a sample survey can be specified as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{2.12}$$

where,

- $\mathbf{y}$ is a $(n \times 1)$ vector of observations
- $\mathbf{X}$ is a $(n \times q)$ matrix of observed covariate values. $\mathbf{X}$ is also referred to as the fixed effect design matrix.
- β is a $(q \times 1)$ vector of coefficients for the covariates (i.e. there are q covariates included in this model or $q - 1$ if a mean response is included.)
- $\mathbf{Z}$ is a $(n \times t)$ matrix of known values indicating the presence of random effects. $\mathbf{Z}$ is also referred to as the random effect design matrix. $\mathbf{Z}$ provides an indication of group memberships.
- $\mathbf{u}$ is a $(t \times 1)$ vector of random effects (i.e. there are t random effects included in the model).
- $\mathbf{e}$ is a $(n \times 1)$ vector of residuals.

If we also assume that both $\mathbf{e}$ and $\mathbf{u}$ are independent and normally distributed with expected values of zero and variances matrices of $\mathbf{R}$ and $\mathbf{D}$ respectively, we can see that the response $\mathbf{y}$ will also be normally distributed with variance matrix

$$\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^{\mathrm{T}} + \mathbf{R} \tag{2.13}$$

where $\mathbf{D} = \mathrm{E}\left[\mathbf{u}.\mathbf{u}^{\mathrm{T}}\right]$ and $\mathbf{R} = \mathrm{E}\left[\mathbf{e}.\mathbf{e}^{\mathrm{T}}\right]$.

Model (2.12) can then be generalized to the GLMM (see McCulloch and Searle, 2001) where the response vector, $\mathbf{y}$, consists of independent elements distributed according to some member of the exponential family, i.e. $y_i|\mathbf{u} \sim$ indep $\mathrm{f}_{\mathrm{y_i}|\mathbf{u}}(\mathrm{y_i}|\mathbf{u})$.

Then a known *link* function, $g(.)$, can be used to relate the mean of the response conditional on the random effects $\mathrm{E}\left[y_i|\mathbf{u}\right] = \mu_i$ to the linear predictor η such that

$$g\left(\mu_i\right) = x_i^T\beta + z_i^T u = \eta$$

Then if $\mathbf{V}$ is known, the unknown parameter β can be estimated by maximum likelihood and the usual weighted least squares estimator $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$ is also the maximum likelihood estimator for β for a normally distributed response vector $\mathbf{y}$ with $\mathbf{V}$ known and provided $\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}$ is non-singular.

In the case where we are interested in estimating the interviewer effect the variance-covariance matrix of response $\mathbf{V}$ is generally not known and its estimation will be the object of the analysis. This adds another $\frac{n}{2}\left(n+1\right)$ unknown parameters to be estimated in the model. By making model assumptions concerning the structure of $\mathbf{V}$ we further reduce the number of parameters we must estimate.

In the following sections we will show how common ANOVA-style multilevel models for the estimation of the interviewer effect can be written as a GLMM. Introducing a basic ANOVA-style model, a survey design structure will be assumed and incorporated into the GLMM via the random effect design matrix, $\mathbf{Z}$ and the fixed effect design matrix, $\mathbf{X}$. The impact on estimation of the interviewer effect in the variance-covariance matrix of response will then be derived.

2.4.3 Example: The Hierarchical Linear Response Model

We will consider the form of the variance-covariance matrix $\mathbf{V}$ in the hierarchical case where respondents, i, are nested within the allocation of respondents to interviewer, j, which are nested within geographical areas k. A simple random response model, with mean response μ, to describe this situation would be

$$y_{ijk} = \mu + Spat_k + Int_{jk} + e_{ijk}$$

where $Spat_k \sim \mathrm{N}\left(0, \sigma^2_{spat}\right)$, $Int_{jk} \sim \mathrm{N}\left(0, \sigma^2_{int}\right)$ and $e_{ijk} \sim \mathrm{N}\left(0, \sigma^2_e\right)$ with $Spat_k$, Int_{jk} and e_{ijk} independent.

Assume for example that we have a balanced hierarchical situation with 8 respondents, 4 interviewers and 2 geographical areas. Then one possible representation of this system in matrix form, of $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$, would be

$$\begin{bmatrix} y_{111} \\ y_{211} \\ y_{321} \\ y_{421} \\ y_{532} \\ y_{632} \\ y_{742} \\ y_{842} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Spat_1 \\ Spat_2 \\ Int_1 \\ Int_2 \\ Int_3 \\ Int_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix}$$

with $\mathbf{R} = \sigma_e^2 \mathbf{I}_8$, where $\mathbf{I}_8$ is the (8×8) identity matrix, and

$$\mathbf{D} = \begin{bmatrix} \sigma_{spat}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{spat}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{int}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{int}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{int}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{int}^2 \end{bmatrix}$$

Then applying (2.13) we can see that the variance-covariance matrix of the response $\mathbf{y}$ will be a block diagonal matrix

$$\mathbf{V} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix}$$

where $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} \sigma_e^2 + \sigma_{spat}^2 + \sigma_{int}^2 & \sigma_{spat}^2 + \sigma_{int}^2 & \sigma_{spat}^2 & \sigma_{spat}^2 \\ \sigma_{spat}^2 + \sigma_{int}^2 & \sigma_e^2 + \sigma_{spat}^2 + \sigma_{int}^2 & \sigma_{spat}^2 & \sigma_{spat}^2 \\ \sigma_{spat}^2 & \sigma_{spat}^2 & \sigma_e^2 + \sigma_{spat}^2 + \sigma_{int}^2 & \sigma_{spat}^2 + \sigma_{int}^2 \\ \sigma_{spat}^2 & \sigma_{spat}^2 & \sigma_{spat}^2 + \sigma_{int}^2 & \sigma_e^2 + \sigma_{spat}^2 + \sigma_{int}^2 \end{bmatrix}$$

which only contains 2 unknown variance component parameters σ^2_{spat} and σ^2_{int} to be estimated and the residual variance σ^2_e. The variance-covariance matrix can then be simplified as the sum of three block diagonal matrices

$$\mathbf{V} = \sigma^2_e \mathbf{I}_8 + \sigma^2_{spat}\mathbf{Z}_2\mathbf{Z}_2^T + \sigma^2_{int}\mathbf{Z}_3\mathbf{Z}_3^T$$

Where $\mathbf{Z}_2$ is an (8×4) matrix containing the columns of $\mathbf{Z}$ that refer to the second, or interviewer, level. Similarly $\mathbf{Z_3}$ will be an 8×2 matrix containing the columns of $\mathbf{Z}$ that identify the groups at the spatial zone level k. Thus in this case we have

$$\mathbf{Z_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{Z_3} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

A more detailed discussion of the $\mathbf{Z}_2$ notation for columns of the random effect design matrix in GLMMs is presented in Chapter 4. Note that the number of rows of $\mathbf{Z_2}$, 8 refers to the number of observations i in the data, while the number of columns, 4, refers to the number of groups at that level, i.e. we have $j = 4$ interviewers.

A further generalization, which would affect the number of parameters to be estimated, is to allow there to be different within group variances at each of the higher levels. This is the approach adopted by Haslett (2003) who considers estimability conditions for disentangling interviewer and small-area effects in his connectivity work. In our example above, this is a relaxation of the earlier assumptions that $Spat_k \sim \mathrm{N}\left(0, \sigma^2_{spat}\right)$ and $Int_{jk} \sim \mathrm{N}\left(0, \sigma^2_{int}\right)$ to allow the variance component introduced by each interviewer and spatial zone to be individually estimated, or in other words $Spat_k \sim \mathrm{N}\left(0, \sigma^2_{spat,k}\right)$ and $Int_{jk} \sim \mathrm{N}\left(0, \sigma^2_{int,j}\right)$. This would mean that instead of there being $l - 1$ extra variance component parameters to estimate in our model,

where l is the number of hierarchical levels in the data, we would need to estimate t extra parameters, where t is the number of random effects included in the model. In our example t would be equal to 6, or the sum of the number of spatial zones and the number of interviewers in our dataset. This would lead to a new $\mathbf{D}$ matrix

$$\mathbf{D} = \begin{bmatrix} \sigma^2_{spat,1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{spat,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{int,1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{int,2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{int,3} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{int,4} \end{bmatrix}$$

and imply a new block diagonal structure to our variance-covariance matrix $\mathbf{V}$ in which there are 6 unknown variance component parameters, $\sigma^2_{spat,1}$, $\sigma^2_{spat,2}$, $\sigma^2_{int,1}$, $\sigma^2_{int,2}$, $\sigma^2_{int,3}$, $\sigma^2_{int,4}$ and the residual variance σ^2_e to be estimated. In this case it can be shown that $\mathbf{V}$ will be

$$\mathbf{V} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}$$

with $\mathbf{A_1}$ equal to

$$\mathbf{A}_1 = \begin{bmatrix} \sigma^2_e + \sigma^2_{spat,1} + \sigma^2_{int,1} & \sigma^2_{spat,1} + \sigma^2_{int,1} & \sigma^2_{spat,1} & \sigma^2_{spat,1} \\ \sigma^2_{spat,1} + \sigma^2_{int,1} & \sigma^2_e + \sigma^2_{spat,1} + \sigma^2_{int,1} & \sigma^2_{spat,1} & \sigma^2_{spat,1} \\ \sigma^2_{spat,1} & \sigma^2_{spat,1} & \sigma^2_e + \sigma^2_{spat,1} + \sigma^2_{int,2} & \sigma^2_{spat,1} + \sigma^2_{int,2} \\ \sigma^2_{spat,1} & \sigma^2_{spat,1} & \sigma^2_{spat,1} + \sigma^2_{int,2} & \sigma^2_e + \sigma^2_{spat,1} + \sigma^2_{int,2} \end{bmatrix}$$

and $\mathbf{A_2}$ equal to

$$\mathbf{A}_2 = \begin{bmatrix} \sigma^2_e + \sigma^2_{spat,2} + \sigma^2_{int,3} & \sigma^2_{spat,2} + \sigma^2_{int,3} & \sigma^2_{spat,2} & \sigma^2_{spat,2} \\ \sigma^2_{spat,2} + \sigma^2_{int,3} & \sigma^2_e + \sigma^2_{spat,2} + \sigma^2_{int,3} & \sigma^2_{spat,2} & \sigma^2_{spat,2} \\ \sigma^2_{spat,2} & \sigma^2_{spat,2} & \sigma^2_e + \sigma^2_{spat,2} + \sigma^2_{int,4} & \sigma^2_{spat,2} + \sigma^2_{int,4} \\ \sigma^2_{spat,2} & \sigma^2_{spat,2} & \sigma^2_{spat,2} + \sigma^2_{int,4} & \sigma^2_e + \sigma^2_{spat,2} + \sigma^2_{int,4} \end{bmatrix}$$

If the aim of specifying separate variance components for each interviewer is to allow the analyst to differentiate between interviewers, the structure described above may be difficult to estimate in practice due to its complexity. An alternative approach demonstrated by Pickery and Loosveldt (2004) allows the identification of exceptional interviewers through examination of the interviewer level residuals. The technique of Pickery and Loosveldt (2004) is a natural extension to the GLMM framework for the estimation of the interviewer effect as it does not require increasingly complex models to differentiate interviewers.

Given that the data has been sorted from the highest level down, the general form of the variance covariance matrix $\mathbf{V}$ can be expressed as follows. In the simple hierarchical case, where there are equal numbers of elements enumerated by each interviewer and in each zone, $\mathbf{V}$ will contain l unknown parameters where l is the number of levels in the data. Then $\mathbf{V}$ will be block diagonal with the following form

$$\mathbf{V} = \sigma_e^2 \mathbf{I}_n + \sigma_{spat}^2 \mathbf{Z}_k \mathbf{Z}_k^T + \sigma_{int}^2 \mathbf{Z}_j \mathbf{Z}_j^T$$

McCulloch and Searle (2001, p 161) present the general case in which the above result is extended to t levels, by adding further variance components. In this case

$$\mathbf{V} = \sum\nolimits_{i=0}^{t} \mathbf{Z_i} \mathbf{Z_i^T} \sigma_i^2$$

where $\sigma_0^2 = \sigma_e^2$ and $\mathbf{Z_0} = \mathbf{I_n}$. Provided the data structure remains hierarchical other generalizations can also be adopted. If unequal numbers of observations are in each unit at a higher level, this will lead to an unequal diagonal structure for $\mathbf{ZZ^T}$. Provided sufficient data, allowing unequal group sizes in the Hierarchial Linear Model (HLM) has little overall effect on estimation of the variance-covariance matrix, as both $\mathbf{ZZ^T}$ and $\mathbf{V}$ will retain a block diagonal structure and the number of unknown parameters to estimate will remain unchanged.

In a similar way further complexity can be introduced into the model. Random effects which are not normally distributed and a variety of response distributions can also be considered, see McCulloch and Searle (2001).

In summary we can say that in the hierarchical linear response model with **V** unknown, we can use knowledge regarding the block diagonal structure of the variance-covariance matrix to reduce the total number of parameters to be estimated in the mixed model from $q+\frac{n}{2}(n+1)$ to $q+l+1$ (this would increase to $q+t+1$ if separate variance components are estimated for each interviewer and spatial area). Furthermore we can see that the hierarchical design of the HLM leads to a block diagonal structure to **V** and hence the interviewer effect will appear separately from the spatial effect in some elements of **V**. Consequently the interviewer effect will never be fully confounded with spatial effects in the simple hierarchical linear response scenario and hence the interviewer effect will be estimable in this case.

2.4.4 Variance-Covariance Matrices in Cross-Classified and Confounded Data Structures

The simplest way to view cross-classified and multiple role data structures is to consider there to be at least two separate hierarchical structures at the same level. In the case of our earlier example this could occur when respondents are nested within interviewer's allocations, so that $i \in g_j$, and respondents are also nested within spatial zones, so that $i \in k$, but interviewers are not nested within the same spatial zones, $g_j \not\subset k$. This situation could occur, for example, when travel costs are low and the survey requires interviewers to travel to several different geographic areas to interview respondents.

The cross-classified and multiple role data structures are generalizations of the general HLM. This generalization does not affect the number of parameters to be estimated, a maximum of $q+t+1$, but does imply that the variance-covariance matrix **V** will not be block diagonal and hence there is no guarantee that the interviewer effect will be estimable under these structures. This loss of information regarding the structure of **V** reduces the number of alternative techniques that can be applied to estimating the parameters of the model and may greatly increase the computing time required during estimation.

In the case of our simple hierarchical structure with 3 unknown parameters introduced earlier, we would have full confounding of the random effects due to the interviewer and the spatial zone if we had only 2 interviewers each enumerating an entire spatial zone, so that $g_j \equiv k$. Then, as the hierarchical structure of the data has been maintained, the variance covariance matrix would still be block diagonal

$$\mathbf{V} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix}$$

However in this confounded case there are only 2 separately estimable parameters in our matrix **A**

$$\mathbf{A} = \begin{bmatrix} \sigma_e^2 + \sigma_C^2 & \sigma_C^2 & \sigma_C^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_e^2 + \sigma_C^2 & \sigma_C^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 & \sigma_e^2 + \sigma_C^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 & \sigma_C^2 & \sigma_e^2 + \sigma_C^2 \end{bmatrix}$$

where $\sigma_C^2 = \sigma_{spat}^2 + \sigma_{int}^2$ and since we only estimate σ_C^2 and never separate either σ_{int}^2 or σ_{spat}^2 we cannot isolate the interviewer effect σ_{int}^2 from the spatial effect σ_{spat}^2.

Thus the general form for the variance covariance matrix **V** in the confounded hierarchical case, in which all random effects are normally distributed and where there are equal numbers of elements collected by each interviewer and in each zone can be seen below. In this case we refer to confounding as occurring when all respondents in a spatial region are enumerated by the same interviewer. Consequently the spatial regions and the interviewer allocations will be equivalent in this case and **V** will contain l unknown parameters with $l-h$ isolable, where l is the total number of hierarchical levels in the data structure and h is the total number of confounded levels.

$$
\begin{aligned}
\mathbf{V} &= \sigma_e^2 \mathbf{I}_n + \sigma_{spat}^2 \mathbf{Z}_j \mathbf{Z}_j^T + \sigma_{int}^2 \mathbf{Z}_k \mathbf{Z}_k^T \\
&= \sigma_e^2 \mathbf{I}_n + \left(\sigma_{spat}^2 + \sigma_{int}^2\right) \mathbf{Z}_j \mathbf{Z}_j^T \\
&= \sigma_e^2 \mathbf{I}_n + \left(\sigma_{spat}^2 + \sigma_{int}^2\right) \mathbf{Z}_k \mathbf{Z}_k^T \\
&= \sigma_e^2 \mathbf{I}_n + \sigma_c^2 \mathbf{Z}_k \mathbf{Z}_k^T \\
&= \sigma_e^2 \mathbf{I}_n + \sigma_c^2 \mathbf{Z}_j \mathbf{Z}_j^T
\end{aligned}
$$

Further generalizations to this model can be incorporated by assuming a functional form for the spatial effect that varies within the region. This is different from the multilevel model which assumes constant correlation between all observations in the same region. For example we could assume that the random effect due to the spatial zone is perhaps some function of distance between observations or location. In this way we are explicitly modelling the spatial correlation in the data. The applicability of spatial modelling techniques to aid in the estimation of the interviewer effect will be considered in more detail in Chapter 6.

2.4.5 Classical Techniques for Isolating the Interviewer Effect

Classical techniques for isolating the interviewer effect in confounded situations have relied upon some form of repeated measurement of each spatial region. This is generally done through adapting the survey design, either so that the survey is interpenetrated (i.e. at least two interviewers are randomly allocated to each spatial region) or so that the same respondent can be re-interviewed by another interviewer.

In the interpenetrated scenario we will have repeated observation involving the spatial effect σ_{spat}^2 with different interviewers and be able to use this information to isolate and estimate the interviewer effect σ_{int}^2 in our variance-covariance matrix $\mathbf{V}$.

In the re-interview situation, if we assume that there is no memory build-up associated with the re-interview process, and for simplicity that a respondent's answer

does not change over time, then we would have repeated observation of the spatial zone at 2 different time points. The re-interview process can be viewed as another hierarchical layer in our data structure where all of the observations collected at time t from a respondent i can be considered to be hierarchically nested within that respondent's set of answers. Thus the response at time t, r_t can be considered to be nested within respondent i, the interviewer at time t, g_{j_t} and the spatial zone k. Thus $r_t \in i \in g_{j_t} \in k$ in the hierarchical situation.

In this case $\mathbf{V}$ may still be block-diagonal overall but we cannot express $\mathbf{V}$ as the sum of a number of block-diagonal matrices. This occurs as we have an overall hierarchical situation (e.g. all responses r_t are nested within spatial zones k) but within the block-diagonal structure imposed by the hierarchy within each zone, we face a cross classified scenario. In other words we are dealing with a multiple-role situation in which each respondent faces at least two different interviewers at each of the time periods, t, so the structure within these larger diagonal blocks will no longer be block diagonal.

Showing this by way of example we present one realization of the re-interview situation. Assume we have a scenario with 2 re-interview periods, 4 respondents, 2 spatial zones and 2 interviewers who interview half the respondents at time $t = 1$ and the remaining half at time $t = 2$. Then the variance-covariance matrix of the response would still have the following block-diagonal structure

$$\mathbf{V} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix}$$

but $\mathbf{A}$ would no longer be block-diagonal

$$\mathbf{A} = \begin{bmatrix} \sigma^2_{spat} + \sigma^2_{int} + \sigma^2_e & \sigma^2_{spat} & \sigma^2_{spat} + \sigma^2_{int} & \sigma^2_{spat} \\ \sigma^2_{spat} & \sigma^2_{spat} + \sigma^2_{int} + \sigma^2_e & \sigma^2_{spat} & \sigma^2_{spat} + \sigma^2_{int} \\ \sigma^2_{spat} + \sigma^2_{int} & \sigma^2_{spat} & \sigma^2_{spat} + \sigma^2_{int} + \sigma^2_e & \sigma^2_{spat} \\ \sigma^2_{spat} & \sigma^2_{spat} + \sigma^2_{int} & \sigma^2_{spat} & \sigma^2_{spat} + \sigma^2_{int} + \sigma^2_e \end{bmatrix}$$

Thus in the re-interview case presented above the repeated measurement has not

changed the number of unknown parameters to be estimated in the model, still σ^2_{spat}, σ^2_{int} and σ^2_e, but it has changed the size and the structure of the variance-covariance matrix of response $\mathbf{V}$ so that σ^2_{spat} and σ^2_{int} are separately estimable.

2.4.6 Summary

We have seen that classic techniques for isolating the interviewer effect, such as interpenetration and re-interviewing, utilize the the methodologist's ability to control aspects of the survey design to re-structure the variance-covariance matrix of response so that the interviewer effect is estimable. Unfortunately, altering aspects of the survey design can prove to be an expensive process and neither interpenetration nor re-interview surveys are commonly applied in practice.

More recently theoretical advances in the estimation of GLMMs and technical advances in computing power have allowed us to fit more complex models and cater more appropriately for the structure of our data. It has also been recognized that this information regarding the hierarchical structure of the data could be used to reduce the number of unknown parameters in the model and enhance our ability to estimate the interviewer effect in hierarchical, cross-classified data structures. This is a passive and relatively inexpensive utilization of information usually available to the survey designer, but still does not allow us to isolate the interviewer effect in situations in which the allocation of respondents to interviewers is fully confounded with spatial zones.

Prior to 1980 estimation of the interviewer effect relied on expensive, active manipulation of the survey design, such as interpenetration and re-interviewing techniques. More recently increases in computing power have allowed passive inclusion of survey design information into multilevel models that facilitated estimation of the interviewer effect in a number of design structures. This thesis will now examine how we can use the remaining spatial and temporal information available to the survey statistician to isolate and estimate the interviewer effect in what has previously been a fully confounded situation.

The following section will look at how variance components such as the interviewer effect can be estimated in practice.

2.5 An Overview of Parameter Estimation in Mixed and Multilevel Models

We have already seen that the effect of including an interviewer in a household survey is to introduce a dependency between responses collected by the same interviewer and that this will lead to an increase in the measurement error component of the variance of estimates derived from the survey. Consequently estimation of the variance component due to intra-interviewer correlation will show the overall effect of the characteristics and actions of the interviewer on the precision of survey estimates. The previous section presented the general structure of the variance-covariance matrix in the GLMM for hierarchical, longitudinal, cross-classified and confounded data structures and examined how many variance component parameters will need to be estimated under each scenario. This section will introduce how the variance component parameters can be estimated given this information.

There are a number of different techniques that can be applied when it comes to estimating variance components in a generalized linear mixed model. In general maximum likelihood techniques (and variations thereof) are now preferred over the more traditional ANOVA options. See for example McCulloch and Searle (2001, p 177) who state

'*We firmly endorse that preference, particularly because, as has already been mentioned, ANOVA methods do not apply satisfactorily to generalized linear mixed models.*'

There are still other options, such as Bayesian estimation, minimum variance estimation (MINVAR) and minimum norm estimation (MINQUE) which may be appropriate under some circumstances. In general there is currently no single estimation method that will be the most appropriate for estimating variance components in all

situations. The following section will derive maximum likelihood estimators for the GLMM under linearity and discuss estimation when the response is not normal.

2.5.1 The Maximum Likelihood Estimator Under Normality

We have already seen that the maximum likelihood estimator for the fixed effect parameters, when $\mathbf{V}$ is known, is $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$. By taking the second derivative of the score function with respect to β^{T} we can also derive the Fisher information

$$I(\beta) = -\mathrm{E}\left[\frac{\partial^2 l}{\partial\beta\partial\beta^{\mathrm{T}}}\right] = -\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}$$

To estimate our t variance components $\sigma^2_{U_1}, \ldots, \sigma^2_{U_t}$ relating to the t independent random effects included in the mixed model, we can re-express the variance-covariance matrix as

$$\mathbf{V} = \sigma^2_e\mathbf{I}_n + \sum\nolimits_{i=1}^{t}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\sigma^2_i$$

so that

$$\frac{\partial V}{\partial\sigma^2_i} = \mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$$

By expressing the variance components in this way we increase the flexibility to allow individual estimation of particular variance components.

By differentiation of the log-likelihood, our score function for any variance component parameter, σ^2_i, of the variance-covariance matrix $\mathbf{V}$ will be in general (see for example Longford, 1993)

$$\frac{\partial l}{\partial\sigma^2_i} = -\frac{1}{2}\left[tr\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma^2_i}\right) - (\mathbf{y}-\mathbf{X}\beta)^{\mathrm{T}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma^2_i}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta)\right]$$

and equating to zero and solving this for σ^2_i we get for each i

$$tr\left(\widehat{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right) = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)^{\mathrm{T}}\widehat{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{V}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)$$

Evaluating this expression will then give us our maximum likelihood estimator for variance component i, $\widehat{\sigma_i^2}$. In practice we cannot estimate $\widehat{\sigma_i^2}$ using this expression unless we also have an estimate of the fixed effects parameter $\widehat{\beta}$ and unfortunately our earlier expression for $\widehat{\beta}$ also requires an estimate of the variance-covariance matrix $\widehat{\mathbf{V}}$. Hence numerically solving these equations becomes an iterative task in which based upon an initial estimate we can generate an improved estimate of both $\widehat{\beta}$ and the variance components $\widehat{\sigma_i^2}$ using the score equations. We can then use these estimates to derive the residual variance $\widehat{\sigma_e^2}$ and the variance-covariance matrix $\widehat{\mathbf{V}}$. The resulting estimate of $\widehat{\mathbf{V}}$ can then be fed back in to produce a better estimate of $\widehat{\beta}$ in the next iteration.

Based upon this score function we can also derive the information matrix for the variance components (for a derivation of the general form of the second derivative of the score function see McCulloch and Searle (2001, p 179))

$$I(\sigma^2)_{\{i,j\}} = -\mathrm{E}\left[\frac{\partial^2 l}{\partial\sigma_i^2\partial\sigma_j^2}\right] = \frac{1}{2}tr\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma_i^2}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma_j^2}\right)$$

which, using the properties of the trace of a matrix will, in this case, evaluate to

$$\begin{aligned} I(\sigma^2)_{\{i,j\}} &= \frac{1}{2}tr\left(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^{\mathrm{T}}\right) \\ &= \frac{1}{2}tr\left\{\mathbf{Z}_i^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_j\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_j\right)^{\mathrm{T}}\right\} \end{aligned}$$

for each row i and column j in a matrix, where $i, j = \{1, \ldots, n\}$ and there are n observations.

This maximum likelihood approach for the estimation of variance components in the normal linear mixed model is an extension of likelihood theory. However these estimators do not, in general, consider the effect of the loss of degrees of freedom which arises from the additional requirement of estimating the fixed effect

parameters in the model. If the appropriate degrees of freedom are not considered, as in the simple maximum likelihood estimator above, the estimate of the variance components $\widehat{\sigma_i^2}$ will be biased by $\frac{-r\sigma_i^2}{n}$ where r is the rank of $\mathbf{X}$. Particularly in the case where this bias will be large, Restricted Maximum Likelihood estimators (REMLs) can be derived to produce a more appropriate estimate of the variance components.

2.5.2 Restricted Maximum Likelihood Estimators

Restricted maximum likelihood estimation relies upon finding a linear combination of the responses that does not contain any of the fixed effects. In other words if we can find some matrix $\mathbf{K}$ such that $\mathbf{K}^{\mathrm{T}}\mathbf{y} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{K}^{\mathrm{T}}\mathbf{V}\mathbf{K}\right)$ then our estimation problem is reduced to just estimating the variance components because our restricted likelihood expression does not depend on the fixed effects parameter β.

Then our restricted maximum likelihood estimators can be derived in the same way as before recognising, that the following transformation has occurred

$$\mathbf{K}^{\mathrm{T}}\mathbf{y} = \mathbf{K}^{\mathrm{T}}\mathbf{X}\beta + \mathbf{K}^{\mathrm{T}}\mathbf{Z}\mathbf{u} + \mathbf{K}^{\mathrm{T}}\mathbf{e} = \mathbf{K}^{\mathrm{T}}\mathbf{Z}\mathbf{u} + \mathbf{K}^{\mathrm{T}}\mathbf{e}$$

and that the variance-covariance matrix of our transformed response will be

$$\mathrm{Var}(\mathbf{K}^{\mathrm{T}}\mathbf{y}) = \mathbf{K}^{\mathrm{T}}\mathbf{V}\mathbf{K}$$

Hence our previous expressions for the maximum likelihood estimators and the REML estimators are equivalent after making the appropriate substitutions for $\mathbf{y}, \mathbf{Z}, \mathbf{X}$ and $\mathbf{V}$ based on the transformation above.

The maximum likelihood estimators we have derived so far have all relied upon the normality of the response vector $\mathbf{y}$. We have already seen that we can generalize our Linear Mixed Model to include situations in which the response is a non-linear combination of fixed and random effects by including a link function $g(.)$. The next section will examine how the simple parameter estimation techniques already discussed can be extended to this more general situation.

2.5.3 Generalized Linear Mixed Models

Starting with the GLMM introduced earlier.

$$g\left(\mathrm{E}\left[y_i|\mathbf{u}\right]\right) = \mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}$$

We will assume a distribution to the random effects $\mathbf{u}$, $\mathbf{u} \sim f_{\mathbf{U}}(\mathbf{u})$ and adopt McCulloch and Searle's notation for the conditional mean and variance of the response y which indicates the dependence of the conditional variance of the response on its conditional mean

$$\begin{aligned}\mathrm{E}\left[y_i|\mathbf{u}\right] &= \mu_i \\ \mathrm{V}\left[y_i|\mathbf{u}\right] &= \tau^2\mathrm{V}\left(\mu_i\right)\end{aligned}$$

Then we can see that the general variance structures derived to consider the interviewer earlier will not necessarily hold under some specifications of the link function. This is because the expressions for the unconditional mean and variance of the response $\mathbf{y}$ cannot be simplified in general. If we first note that a simple rearrangement of our definition of the link function will provide

$$\mu_i = g^{-1}\left(\mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}\right)$$

then the unconditional mean of the response in the GLMM will be

$$\begin{aligned}\mathrm{E}\left[y_i\right] &= \mathrm{E}\left[\mathrm{E}\left[y_i|\mathbf{u}\right]\right] \\ &= \mathrm{E}\left[g^{-1}\left(\mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}\right)\right]\end{aligned}$$

which we can only simplify, as before, if the link function is linear. Otherwise this general expression must be evaluated for each specific link function g.

Similarly, under a non-linear link function the variance-covariance matrix of the response $\mathbf{y}$ must be evaluated for each specific link function g with elements

$$
\begin{aligned}
\mathrm{cov}\left(y_i, y_j\right) &= \mathrm{cov}\left(\mathrm{E}\left[y_i|\mathbf{u}\right], \mathrm{E}\left[y_j|\mathbf{u}\right]\right) + \mathrm{E}\left[\mathrm{cov}\left(y_i, y_j|\mathbf{u}\right)\right] \\
&= \begin{cases} \mathrm{v}\left(g^{-1}\left(\mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}\right)\right) + \mathrm{E}\left[\tau^2 \mathrm{v}\left(g^{-1}\left(\mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}\right)\right)\right] & \text{if } i = j \\ \mathrm{cov}\left(g^{-1}\left(\mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}\right), g^{-1}\left(\mathbf{x}_j^{\mathrm{T}}\beta + \mathbf{z}_j^{\mathrm{T}}\mathbf{u}\right)\right) & \text{otherwise} \end{cases}
\end{aligned}
$$

for each row i and column j in the variance-covariance matrix of response $\mathbf{V}$. Note that although this expression cannot generally be evaluated without making further assumptions to determine the form of the link function g, the covariance term will still be zero between observations which do not have any random effects in common. Hence in the GLMM the number of parameters requiring estimation and the general structure of the variance-covariance matrix of response $\mathbf{V}$ will not have changed for the hierarchical, cross-classified, longitudinal and confounded scenarios discussed earlier even though the form of the individual elements in $\mathbf{V}$ may have changed.

Given that we know the form of the link function g we can write down the distribution of the response $\mathbf{y}$ conditional on the random effects $\mathbf{u}$, or $f_{Y_i|\mathbf{u}}\left(y_i|\mathbf{u}\right)$. Using this expression we can write down the general form for the conditional likelihood function

$$
l\left(\beta, \varphi|\mathbf{u}\right) = \prod_{i=1}^{n} f\left(y_i|\mathbf{u}\right)
$$

where β is the vector of fixed effect parameters and φ is the vector random effect parameters. We can then use this to obtain the unconditional likelihood function by integrating out the random effects

$$
l\left(\beta, \varphi\right) = \int \prod_{i=1}^{n} f\left(y_i|\mathbf{u}\right) f\left(\mathbf{u}\right) d\mathbf{u}
$$

Unfortunately this expression is often intractable and maximization may require the use of numerical integration techniques. McCulloch and Searle (2001, p 227)

present some further simplifications of the score function based upon this likelihood expression and discuss numerical techniques such as the EM algorithm, the Newton-Raphson method, numerical quadrature, Markov Chain Monte Carlo methods, simulation techniques and stochastic approximation algorithms. There are a number of competing methods for estimating parameters in GLMMs and McCulloch and Searle (2001) suggest that none of these techniques has thus far been shown to be uniformly best for all subclasses of the GLMM. Other approaches, including the hierarchical likelihood estimators of Lee and Nelder (1996), are still being explored, but determination of an appropriate estimation procedure will generally be dependent on the dataset to be analyzed.

A number of software packages include options for fitting multilevel models. The next section will examine how MLwiN approaches the estimation of parameters in the GLMM.

2.5.4 Multilevel Model Parameter Estimation in MLwiN

MLwiN is the latest variant of the specialized multilevel modelling package of which earlier versions were ML3 and MLn. MLwiN utilizes the IGLS and the RIGLS algorithms developed in Goldstein (1986, 1989) while later versions also include a MCMC engine developed by Browne (1998). The IGLS and RIGLS algorithms work separately on two different blocks of parameters. First estimates for the fixed effects parameters β are generated given an initial estimate of the variance-covariance matrix of response $\mathbf{V}$, say $\tilde{\mathbf{V}}$, by using

$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}\mathbf{y}$$

and then these estimates of the fixed effects parameters are used to estimate the variance component parameters assuming that the initial estimates of β are fixed. Estimation then proceeds iteratively until both parameter estimates stabilize.

More detail on the estimation of the variance components given our estimate of β can be found in Goldstein and Rasbash (1992). In summary given $\hat{\beta}$ the vector of

'raw' residuals $\tilde{\mathbf{y}}$ is formed from

$$\tilde{y}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_{ij}$$

and then our estimate of the variance-covariance matrix of response is derived by recognizing that

$$\hat{\mathbf{V}} = \mathrm{E}\left[\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathrm{T}}\right]$$

Both of these expressions can be solved using generalized least squares and hence the name IGLS. Note also that although this procedure is applied in MLwiN there is a perceived lack of documentation. For example de Leeuw and Kreft (2001, p 191) write

'*It is not entirely clear from the documentation what happens in boundary cases where dispersion matrices become singular or even indefinite.*'

On the other hand there is clear documentation of the MCMC engine available in MLwiN version 1.20 and higher. The MCMC engine in MLwiN was initially developed by Browne (1998) and as a default adopts an adaptive Metropolis-Hastings sampler. The following section briefly outlines the Bayesian estimation approach for parameters in GLMMs.

2.6 Bayesian Estimation of Parameters in Multilevel Models

Bayesian statistics have become more widely utilized as availability of powerful computers has increased. Bayesian applications that once would have been impossible to implement in practice can now be implemented with reasonable efficiency. One such application is the estimation of parameters of multilevel models. This section will summarize briefly how Bayesian techniques can be used to estimate parameters in multilevel models and the advantages/disadvantages of these techniques in

comparison with the more traditional frequentist approach highlighted in previous chapters.

2.6.1 Bayesian Inference

Rather than considering the vector of fixed and random effect parameters θ to be fixed, as in the frequentist approach, Bayesian inference assumes that θ is random and attempts to explore its distribution given the observed data, y. The conditional distribution of θ can then be explored via the following application of Bayes' theorem (see for example Gilks *et al.*, 1996)

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} \tag{2.14}$$

Hence if we can express the current state of our knowledge regarding the parameter vector θ as a prior distribution, $P(\theta)$, and combine this with information regarding the observed data, y, we can produce a (conditional) posterior distribution, $P(\theta|y)$, from which we can estimate characteristics and functional summaries of the parameter vector based on the observed data.

In high dimensional problems the integrals in (2.14) may not have a closed form solution for the posterior distribution (Breslow and Clayton, 1993). However even in lower dimensional scenarios evaluating these integrals has traditionally been the stumbling block to efficient application of Bayesian inference. As analytic evaluation of (2.14) may be impossible a number of alternative approaches have been considered, such as numerical evaluation, analytic approximation (e.g. Kass *et al.*, 1988) and Monte Carlo integration.

2.6.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) simulation techniques (Gelfand and Smith, 1990) have become increasingly popular over the last decade. MCMC techniques are a flexible methodology that can be easily adapted for use with a wide vari-

ety of models. Hence software packages that utilise empirical Bayes techniques for fitting multilevel models, such as MLwiN 2.0 (Rasbash *et al.*, 2003) and BUGS (Spiegelhalter *et al.*, 1994) both use general purpose MCMC samplers as their fitting mechanisms.

MCMC methods work by generating samples from Markov chains which converge to the posterior distribution of interest, $P(\theta|y)$. Upon successful convergence of the MCMC chain we therefore have a sample from the posterior distribution, from which we can evaluate summaries of the parameter vector, avoiding the need to evaluate potentially intractable integrals in (2.14).

The next section will summarize MCMC techniques for estimation of multilevel model parameters in practice.

Monte Carlo Integration and Markov Chains

Monte Carlo integration can be used to evaluate summary statistics by directly sampling from the distribution of interest. When these samples are independent the accuracy of the summaries is regulated by laws of large numbers, but drawing independent samples is often not feasible. This lack of independence will not affect the results provided the samples span the distribution and are drawn in the correct proportions. This can be achieved by setting up a Markov chain with the appropriate stationary distribution (see Roberts (1996); Tierney (1996) and Gilks *et al.* (1996) for more information).

The Metropolis Hastings Algorithm

The Metropolis algorithm (Metropolis *et al.*, 1953) is a simple method for constructing a Markov chain with an appropriate stationary distribution. After constructing a proposal distribution which has the posterior as its stationary distribution, the algorithm works by correcting draws from this proposal distribution so that we are actually simulating from the posterior, $P(\theta|y)$. The way in which this is done is to only accept new values drawn from the proposal distribution in each step if they

pass a related test criterion. Hence each time a proposed value is accepted the estimates of the parameters of interest are improved as the Markov chain approaches its stationary distribution - the posterior in which we are interested.

It is important to note that the choice of the proposal distribution will affect the performance of the Metropolis algorithm. In particular the value given to the variance parameter of the proposal will affect how well the simulation performs. If the variance of the proposal is too small the proposal will not vary much from its mean and consequently the algorithm may take a long time to converge as each step will be small. In contrast if the variance of the proposal distribution is too large we may reject a large proportion of the proposed values again, slowing down convergence. Adaptive techniques are sometimes used to improve convergence, by modifying the parameters of the proposal distribution based on the characteristics of the data (see Browne (1998) for a discussion of the adaptive process adopted in MLwiN).

The Metropolis-Hastings algorithm (Hastings, 1970) is a generalization of the Metropolis algorithm that allows for proposal distributions that are not symmetric.

Gibbs Sampling

In practice the Metropolis-Hastings algorithm can often be simplified, especially when the posterior can be specified in terms of its full conditional distributions. It may be easier to sample from the conditional distributions instead of the marginal posterior and if this is the case Gibbs sampling can be used to sample indirectly from the posterior. Gibbs sampling was described by Geman and Geman (1984) in the field of image analysis, before being applied to statistical problems by Gelfand and Smith (1990). As it only samples from full conditional distributions, Gibbs sampling can be considered a special case of the Metropolis-Hastings algorithm.

2.6.3 MCMC in Practice

MCMC methods are among the most flexible techniques available for conducting Bayesian inference. As such software packages that utilize empirical Bayes techniques for fitting multilevel models commonly use general purpose MCMC samplers for their fitting mechanisms. For example BUGS (Spiegelhalter *et al.*, 1994) utilizes Gibbs sampling with the adaptive rejection algorithm of Gilks and Wild (1992), while MLwiN 2.0 (Rasbash *et al.*, 2003) uses a tailored adaptive Metropolis-Hastings algorithm (Browne, 2002). Since Gibbs sampling can be considered a subset of the Metropolis-Hastings algorithm there appears little to choose between these two techniques, although as Browne (1998, p 43) points out

'*...most of the research in the use of MCMC models has concentrated on Gibbs sampling. This is primarily because of its ease of programming.*'

Browne and Draper (2003) show that the adaptive Metropolis-Hastings algorithm generally outperforms Gibbs sampling based on a study involving a number of models.

This section will examine a number of the practical implementation issues involved with estimation of multilevel model parameters using MCMC methods.

Convergence

In Maximum Likelihood Estimation (MLE), convergence is generally determined by some form of convergence criterion and overall performance assessed by a number of diagnostic tools. With MCMC we are interested in knowing when the chain has appropriately converged to a distribution, rather than an estimate, before using this posterior to make inference regarding the parameters of interest. In practice then, how do we know when the MCMC chain has moved from the starting values and is sampling from the appropriate stationary posterior distribution?

Without indicating whether the MCMC chain is approaching the appropriate stationary distribution, a number of convergence diagnostics have been designed to indicate whether a MCMC chain has converged. There are a large number of po-

tential options such as the Raferty and Lewis (1992) and Geweke (1992) diagnostics applied in MLwiN 2.0 and comprehensive reviews of a number of MCMC convergence diagnostics can be found in Cowles and Carlin (1996) and Brooks and Roberts (1998).

Note that even if the MCMC chain has converged we may be dealing with a multi-modal distribution. If this is the case it is possible that the chain is simulating from a single mode and not the entire distribution. In order to confirm that we do not have a multi-modal distribution the MCMC simulation must be started from a number of different starting values. If the distribution is uni-modal then all chains should eventually converge to produce the same estimates regardless of the starting values, while if the distribution is multi-modal we would see widely differing estimates from chains with different starting values. There are a number of convergence diagnostics that rely on multiple chains with different starting values in order to consider this issue; see for example Gelman and Rubin (1992).

Length of MCMC Run

Once convergence has been achieved, and through this an appropriate burn-in length determined, it remains to decide how large a sample we need to draw from our posterior to produce estimates of a required accuracy. The number of iterations (and thereby the length of time) over which we need to run the chain is not always clear as subsequent iterations may not be independent. As Browne (1998, p 38) recognizes

'*Auto-correlation is an important issue when considering the chain length, as a chain that is mixing badly, that is, has a high auto-correlation will need to be run longer to give estimates of a required accuracy.*'

Exploration of the auto-correlation function (ACF) and the partial autocorrelation function (PACF) will generally indicate the presence of auto-correlation and the Rafferty-Lewis diagnostic can be used to create a comparison of the relative efficiency of a given MCMC chain against independence.

Provided convergence has been achieved, auto-correlation will not inhibit the

MCMC chain sampling the entire posterior and producing appropriate estimates. However the length of run required to explore this distribution may be greatly increased. If autocorrelation is a problem the following techniques can be applied to improve mixing and reduce required chain lengths.

- Thinning. A technique that stores only every kth iteration of the chain. This technique offers only slight speed gains but has the added attractions of reduced storage requirements and less auto-correlation in the thinned chain. MacEachern and Berliner (1994) showed that a thinned chain does not produce as accurate estimates as the complete chain.

- Altering the characteristics of the proposal distribution (especially when adaptive Metropolis-Hastings is not applied) may improve the mixing properties of the chain.

- Re-parametrization (Hill and Smith, 1992) and hierarchical centring (Gelfand *et al.*, 1995) both alter the form of the model being fitted to improve mixing. A practical application of these techniques can be found in Browne (2004)

Discussion

One of the major difficulties with MCMC estimation is knowing when the chain has converged and whether we have enough iterations to make appropriate inference. Most of these problems are related to the choice of starting values, prior and proposal distributions. In practice if we use diffuse priors, an adaptive algorithm and try several different starting values, MCMC methods can be considered to be very flexible and appropriate techniques for the estimation of parameters in multilevel models.

There are a number of issues to be considered with MCMC techniques for the estimation of parameters in multilevel models. In particular model selection criteria under a Bayesian framework have still to be comprehensively defined. Spiegelhalter *et al.* (2002) proposes a Deviance Information Criterion (DIC) to allow model

comparison. However due to time constraints and a lack of consensus on use of the DIC, running a comprehensive Bayesian model search is not always feasible.

2.7 Summary

We have seen that the interviewer effect is related to the correlation between responses collected by the same interviewer. In this chapter we demonstrated that if we do not appropriately consider the interviewer effect this leads to underestimates of the variance of results derived from surveys, potentially leading to inappropriate application of survey data. We have introduced results which can be used to estimate the total variance of estimates derived from interviewer-enumerated surveys. These results will allow us to produce variance estimates of the mean response in survey data that appropriately cater for the interviewer effect for the first time. We have discussed how considering the interviewer effect requires us to estimate the correlation between responses due to the presence and characteristics of the interviewer. Classic techniques for isolating the interviewer effect, such as interpenetration and re-interviewing, utilize the methodologist's ability to control aspects of the survey design to re-structure the variance-covariance matrix of response so that the interviewer effect is estimable. Unfortunately, altering aspects of the survey design can prove to be an expensive process and neither interpenetration nor re-interview surveys are commonly applied in practice.

We have shown how the generalized linear mixed model can be applied to estimate the interviewer effect in complex survey data. The advantage of this GLMM formulation is that it explicitly incorporates survey design information into the model allowing us to comprehensively evaluate the impact of the survey design on estimation of the interviewer effect for the first time. We have discussed how information regarding the hierarchical structure of the data can be utilized to enhance our ability to estimate the interviewer effect in hierarchical, cross-classified and multiple-membership data structures. This is a passive and relatively inexpensive utilization of information usually available to the survey designer, but still does

not allow us to isolate the interviewer effect in situations in which interviewers are not fully interpenetrated within spatial zones.

Previous techniques for estimating the interviewer effect have concentrated on active and passive use of the survey design. The next step in this process is to utilize the remaining spatial, temporal and design information available to the survey statistician to isolate and estimate the interviewer effect in what has previously been considered a fully confounded situation.

Chapter 3

Longitudinal Modelling for Interviewer Effects

In this chapter we will consider how the interviewer effect can be estimated by incorporating already available longitudinal information into variance decomposition models for surveys which are not fully interpenetrated. Some authors have considered how longitudinal information can aid estimation of interviewer effects, such as re-interviews (Felligi, 1964) and over successively interpenetrated waves of a spatially concentrated survey (Pickery and Loosveldt, 2000, 2001). However these studies have relied upon interpenetration and are very expensive to adapt to large-scale surveys, which are generally run on a tight budget. Interpenetration requires random allocation of interviewers to spatial regions and consequently the increase in travel costs required to produce a fully interpenetrated design will generally be prohibitive.

We will show how, if interviewers are rotated over time, longitudinal information can be incorporated as an extra hierarchical level in the classificatory structure of the data. This technique will lead to a large increase in the effective degree of interpenetration on our dataset and hence aid estimation of the interviewer effect. The applicability of these techniques will be demonstrated using confidentialised Aus-

tralian Bureau of Statistics (ABS) Labour Force Survey (LFS) data. This dataset contains primarily objective data items for which the interviewer effect may be small. However based on this data we can demonstrate techniques and develop methodologies for the estimation of the interviewer effect in non-fully interpenetrated surveys.

The following chapters will then consider in general the appropriateness of interviewer effect estimates derived from surveys that are not fully interpenetrated, prepare a more comprehensive definition of interpenetration, present cost-optimal designs for the estimation of the interviewer effect and consider models that describe spatial autocorrelation in the dataset.

3.1 CURF Dataset

This section presents general information on the ABS dataset before extending current multilevel analysis techniques to estimate the interviewer effect by incorporating the longitudinal information available in this dataset.

The dataset used for this study is a Confidentialised Unit Record File (CURF) sample of 50 workloads drawn from the LFS component of the Monthly Population Survey (MPS) over the months August to November 2001. The workloads in the CURF are a collection of households to be enumerated during the LFS and households within a workload are generally in closer geographic proximity than households in different workloads. Consequently the workload can also be viewed as a spatial region, with households enumerated within the workload belonging to this region. Chapter 6 will consider applications of Spatial modelling, which relies on the geographic proximity of units, to estimate the interviewer effect.

The LFS is a large repeated panel household survey, which includes a multistage area sample of private dwellings (houses, flats etc) and a sample of non-private dwellings (hotels, motels, etc) selected in Collection Districts (CDs) from a frame maintained by the ABS. The LFS samples approximately 0.5% of the Australian population, corresponding to around 30,000 households and 65,000 individual respondents each month and has been run since February 1978. The LFS is structured

as an *in-for-8* repeated panel survey, so that each household is selected to remain in the sample for 8 consecutive months, with one eighth of the households being replaced each month. Data is collected from respondents via face-to-face Any Responsible Adult (ARA) interviews in the first month, while subsequent interviews are conducted over the telephone, provided this is acceptable. Data is collected about all respondents in a household from the ARA.

In order to maintain the privacy of personal information collected in the LFS, a confidentialised sample of 50 workloads was extracted. As part of the confidentialisation process all spatial identifiers were removed. The classification structure of the data is still available through group level identifiers, indicating to which household, CD, interviewer or workload an observation belongs. These identifiers have been confidentialised so that no information regarding the relative position of these groups remains in the CURF dataset. In comparison the longitudinal information available in the LFS was not considered a disclosure risk and is retained in the CURF. A number of further adjustments were made to reduce the potential disclosure risk from the CURF, for example households containing more than six responding adults were removed and the weights on the dataset were also adjusted so that regions could not be identified by their size. This chapter will focus on how the remaining longitudinal information in the CURF can be incorporated under simple variance decomposition models to produce more appropriate estimates of the interviewer effect.

3.1.1 Data Issues

This section will examine the structure and content of the CURF dataset. The CURF contains 20 data items,

1. An 8 digit unique identifier comprising three components,

   (a) A randomly generated number identifying the Primary Sampling Unit (PSU), dwelling and household. First 5 digits.

(b) Person identifier within household. Next 2 digits.

(c) Month number. Last 1 digit.

2. Age. Categories 1-6.

   (a) Category 1: 15-24 year olds inclusive
   (b) Category 2: 25-34 year olds inclusive
   (c) Category 3: 35-44 year olds inclusive
   (d) Category 4: 45-54 year olds inclusive
   (e) Category 5: 55-64 year olds inclusive
   (f) Category 6: 65 and above year olds

3. Sex. 1 Male, 2 Female.

4. Marital Status. 1 Lives with partner, 2 Does not live with partner.

5. Country of Birth. 1 born in Australia, 2 born in other English speaking country, 3 otherwise.

6. Labour force status. 1 Employed, 2 Unemployed, 3 Not in the labour force.

7. Full-time, part-time employment status. 0 Not employed, 1 Works full-time, 2 Works part-time.

8. Hours worked. Categories 0-8.

   (a) Category 0: zero hours worked
   (b) Category 1: 1-15 hours worked inclusive
   (c) Category 2: 16-25 hours worked inclusive
   (d) Category 3: 26-30 hours worked inclusive
   (e) Category 4: 31-35 hours worked inclusive
   (f) Category 5: 36-40 hours worked inclusive
   (g) Category 6: 41-45 hours worked inclusive
   (h) Category 7: 46-50 hours worked inclusive
   (i) Category 8: more than 50 hours worked

9. Workload identification number. Randomly generated 5 digit number.
10. Collection District identification number. Randomly generated 5 digit number.
11. Interviewer identification number. Randomly generated 5 digit number.
12. Month. Label 1-4 for August to November 2001.
13. Response Status. 1 Full, 2 Part, 3 Sample loss, 4 Non-response.
14. Partial non-response. 0 All in dwelling in scope, 1 otherwise.
15. Rotation Group. Groups 1 to 8 and missing.
16. Response type. 0 Self enumerated, 1 Personal interview, 2 Any responsible adult enumeration.
17. Interview type. 0 Self enumerated, 1 Telephone interview 2 Face-to-face interview.
18. Total number of clusters in the Collection District.
19. Total number of clusters in the block.
20. Weight for response. Note some weight swapping was performed to protect confidentiality

The LFS is the primary vehicle through which unemployment data is collected in Australia. As such the data item of primary importance in the LFS is labour force status; a data item derived from an individual's responses to approximately 80 objective questions. Due to the importance of labour force status the focus of this chapter will be on estimating the interviewer effect associated with this data item.

The CURF sample contains records obtained from 7,229 individuals, living in 3,643 households, 388 collection districts (CDs), interviewed by 110 different interviewers over 4 different months, giving us a total of 20,227 observations (or rows of data). There are some missing values (recorded as NAs) in the CURF for which closer observation shows these 1,415 observations are missing information on *all* of the following 8 data items,

1. Age

2. Sex
3. Marital status
4. Labour force status
5. Full-time, part-time employment status
6. Response type
7. Interview type
8. Weight for response

All of these 1,415 observations have response status 3 or 4, which identifies them as sample loss or non-response households. All sample loss and non-response households have then been allocated the same interviewer identification number that is not associated with any valid responses. It appears that all sample loss and non-response households have been allocated an identical dummy interviewer identifier and we are not given information regarding which interviewers attempted to contact the households that were later categorized as non-response or sample loss.

As these 1,415 observations do not give us any useful information regarding the influence of the interviewer on non-response (i.e. we are not told which interviewer was allocated to contact the non-responding household) it was decided to discard this data and concentrate on the effect of the interviewer on responses in responding households.

After the records with missing values were excluded the remaining dataset contains

- 18,812 observations or rows of data
- Recorded in 1 of 4 different months
- By 1 of 109 different interviewers
- There are 6,867 individual observations. Note that some of these individuals appear to not be successfully matched across time
- There are 3,308 different households

- The observations have been collected from 50 different workloads, and
- 387 different collection districts

Although an unique individual respondent identifier is available on the CURF it may be difficult to match respondents living within households over time. This can occur, for example, if people move into or out of a household, the ARA forgets to record a number of individuals or due to data entry or recording problems. Hence individuals within a private household may be incorrectly matched from month to month on the CURF and the person level identifier may be misleading.

Partial validation of the person level identifier can be provided by matching an individual's demographic characteristics from month to month. For example, if a person has been correctly matched from one month to the next, their gender should not change, their country of birth will not change and their age cannot decrease. The following list summarizes the conditions that have been adopted in an algorithm to match persons within a household by demographic characteristics

- Age category A_t at time t must be either equal to or 1 less than age categories A_{t+1} , A_{t+2} , A_{t+3} , at time $t+1$, $t+2$ or $t+3$
- Country of birth category cannot change over time
- Gender cannot change over time

Running this algorithm on the CURF indicates that a total of 67 households contain at least one mismatched individual.

In the ABS LFS data comprehensive logical consistency checks are performed for each month individually, but not over time. This algorithm allows us to identify households within which there is some form of measurement error in any individual month which leads to an inconsistency in the data over time. Based on the confidentialised and aggregate data items that are available in the CURF we cannot in general make statements regarding the cause of these errors. For example we may not be able to distinguish a measurement error in one month for the individual respondent identifier from a measurement error in the age of the respondent. An

example of this can be seen in Table 3.1 in which an excerpt from the CURF is presented showing that the individual respondent identifier within a single household appears to have been swapped in October 2001 between the two individuals within the household when compared to the remaining months. Closer examination of the unconfidentialised LFS dataset (which allows matching of non-aggregate demographic variables such as age recorded to the nearest year) within the ABS office later, indicated that this was actually a measurement error in the gender of the respondents rather than a measurement error in the person identifier. Note that Table 3.1 is presented as an example only and consequently the household (HH) and interviewer identifiers presented in Table 3.1 are not the original identifiers that appeared in the CURF.

HH id	Interviewer id	Month	Person id in HH	Gender	Age group
1	1	Aug 2001	1	1	5
1	2	Sep 2001	1	1	5
1	2	Oct 2001	1	2	5
1	2	Nov 2001	1	1	5
1	1	Aug 2001	2	2	5
1	2	Sep 2001	2	2	5
1	2	Oct 2001	2	1	5
1	2	Nov 2001	2	2	5

Table 3.1: Example Drawn from CURF of Measurement Error over Time within Household

We can see from the example presented in Table 3.1 that there are a number of measurement errors in the CURF that make it difficult to match all individuals over time. Measurement errors such as these can generally be identified via probity checks that evaluate the logical consistency of the data over time. Multi-level modelling requires a within household identifier, and given that there is uncertainty regarding the individual identifier for only 2% of the data, we will initially examine the data by excluding households containing apparently inconsistent data over time.

We have defined the interviewer effect as the impact on estimates due to the correlation of responses caused by the presence of the interviewer. Hence the interviewer

effect we are measuring is generally an effect above and beyond the measurement errors highlighted here. However, interviewers that make systematic measurement errors in the recording of the data item of interest will still influence the magnitude of the estimated interviewer effect, even if the majority of the logically inconsistent measurement errors have not first been removed by probity checking. Although this demonstrates the need for more comprehensive probity checking of the consistency of the data over time, estimates of the interviewer effect would otherwise pick up interviewers who make these sort of systematic errors.

After NAs and inconsistent records have been removed from the CURF, the remaining dataset contains

- 18,264 observations or rows of data
- Recorded in 1 of 4 different months
- By 1 of 109 different interviewers
- There are 6,702 individual respondents
- There are 3,241 different households
- The observations have been collected from 50 different workloads
- And 387 different collection districts

In this section we have summarized the data items that are available on the CURF and removed records containing missing values and inconsistent demographic information over time. These records were removed as there was insufficient information remaining in the confidentialised dataset to indicate interviewer effects and to clarify the classification structure of the data. The following section will examine the structure of the resulting dataset in more detail.

3.1.2 Examination of CURF

This section will explore the data structure of the CURF and examine whether we have some form of repeated measurement of interviewers that will allow us to estimate the interviewer effect.

A breakdown of the number of workloads in which an interviewer collects data can be seen in Table 3.2 below

Workloads	Number of Interviewers	% of Total
1	95	87.2
2	12	11.0
3	1	0.9
4	0	0
5	0	0
6	0	0
7	0	0
8	1	0.9
Total	109	100

Table 3.2: Distribution of Number of Workloads from which Interviewers Collect Data. August to November 2001.

We can see that most interviewers collect data from only one workload, although some are involved in the enumeration of several workloads, including one interviewer who is involved in the enumeration of 8 different workloads. No other interviewer collects data from more than three different workloads so it is likely this interviewer has received further training to deal with *unusual* dwellings. These dwellings may have different characteristics to the majority of residential dwellings and which we would expect to contain individuals that will not generally respond in the same way as individuals in the majority of dwellings. Due to this difference, interviewers may receive further training on how to deal with these dwellings. For example an interviewer might have received training for interviewing respondents who live in apartment blocks and hence this interviewer might be required to travel to a large number of workloads in order to specifically deal with respondents living in apartments. Overall the mean number of workloads from which each interviewer collects data is 1.19.

Interviewers also collect data from a mean of 56.5 households and 8.6 CDs, the average CD containing 6.6 sample households. The minimum number of households from which a single interviewer collected data is 1 and the minimum number of CDs is 1. The maximum number of households from which a single interviewer collected

data is 136 and the maximum number of CDs 21.

The information interviewers collect from households relates to a mean of 114.9 respondents per interviewer, with the minimum number of respondents for any interviewer 1 and the maximum 300.

In a similar way we can look at the breakdown by workload to assess the hierarchical structure of the data and determine if there are any workloads with unusual characteristics in the data.

Interviewers in Workload	Number of Workloads	% of Total
1	6	12
2	22	44
3	12	24
4	6	12
5	4	8
Total	50	100

Table 3.3: Number of Interviewers Appearing in each Workload over the 4 Month Period

In the CURF, workloads vary greatly in size, from a minimum of 15 respondents to a maximum of 225 with a mean of 134. We can see in Table 3.3 that the minimum number of interviewers in a workload is 1, while the maximum is 5 with the mean number of interviewers per workload 2.6. Across all workloads there is a minimum of 2 CDs and a maximum of 15 with a mean of 7.7. There is a minimum of 8 households in a workload and a maximum of 95 with a mean of 64.8.

More importantly what Table 3.3 tells us is that if we consider the CURF as a longitudinal dataset, over the entire 4 month period available to us, 88% of the workloads are enumerated by more than 1 interviewer. This occurs due to a policy of enumerating respondents with different interviewers over time and gives us a high degree of repeated measurement (or alternatively a high effective degree of interpenetration) of both workloads and interviewers. This can be compared with Table 3.2 in which over 87% of the interviewers do *not* appear in more than 1 workload. This indicates that we do not have a high degree of repeated measurement of either interviewers or workloads when we consider each month of the CURF

individually and this will have implications regarding how well we can estimate the interviewer effect without using the longitudinal information available in the CURF.

The LFS is a large-scale survey run on a tight budget and consequently it is not feasible to produce a fully interpenetrated design for any single month of the LFS without a subsequent increase in costs. However, although the LFS has not been designed as an interpenetrated survey, we can consider it to have a form of temporal interpenetration, that is we have repeated measurement of the majority of workloads by different interviewers over time. In the following sections we will consider how we can utilize this information to produce an estimate of the interviewer effect.

Considering the CURF without Longitudinal Information

Before considering whether we can use temporal interpenetration to separate the workload level and interviewer effects we should also consider whether we can produce estimates of the interviewer effect by considering each month in isolation. The following table summarizes the number of interviewers, respondents, CDs, households and workloads that are in each month of the CURF.

Month	Respondents	Interviewers	CDs	Hholds	Wklds
Aug 2001	4445	49	383	2245	50
Sep 2001	4512	53	383	2259	50
Oct 2001	4639	50	385	2294	50
Nov 2001	4668	53	383	2311	50

Table 3.4: Summary of CURF Characteristics by Month

We can see that the total number of respondents, interviewers, CDs and households is relatively stable in each of the months, although the individual respondents, interviewers, CDs and households do change slightly from month to month while each of the workloads is enumerated in every month. Less than half of the interviewers are active in any single month between August and November 2001 and this may have implications regarding how many interviewers can be matched over time and whether we could potentially estimate separate interviewer effects for each interviewer. For an algorithm that can be used to determine whether there is enough

information available in the CURF for the estimation of separate interviewer effects (i.e. separate variance components for each interviewer), see the connectivity work of Haslett (2003). In Chapter 1 we defined the interviewer effect we are interested in measuring based on the traditional definition that the interviewer effect is the impact on estimates derived from the survey caused by the unintended correlation between response due to the presence of the interviewer. Consequently we are not aiming to estimate separate variance components for each interviewer. Furthermore we can distinguish between interviewers under a GLMM variance decomposition model by considering the interviewer level residuals (see for example Pickery and Loosveldt (2004)). To consider whether we can estimate separate variance components for each interviewer, Table 3.5 looks at how many interviewers, respondents, households and CDs can be matched in the data for 1 month, 2, 3 or all 4 months.

	1 Month	2 Months	3 Months	4 Months	Total
Interviewers	48	36	15	10	109
Respondents	1560	1269	1326	2547	6702
Households	663	604	658	1316	3241
C. Districts	4	0	2	381	387

Table 3.5: Number of Months in CURF for Interviewers, Respondents, Households and CDs

We can see that 44% (or 48 out of 109) of the interviewers are in the sample for only one month out of the four and that less than half of the respondents and households are in all four months. As a repeated panel survey, it was expected that approximately half of the respondents and households would be rotated out of the LFS over a four month period.

Table 3.6 summarizes how many interviewers are in each workload in each month. Given that we have already examined the total number of different interviewers in each workload over all four months (see Table 3.3), it is clear that interviewers are swapped over time in most workloads.

The next question, when assessing the classificatory hierarchy is whether an interviewer can appear in more than one CD in a single month. A summary table

Interviewers in Workload	Mth 1	Mth 2	Mth 3	Mth 4
1	45	43	45	40
2	4	6	5	9
3	1	1	0	1
Total	50	50	50	50

Table 3.6: Number of Workloads from which Interviewers Collect Data, by Interviewer by Month

for the number of interviewers in CDs in the CURF dataset can be seen in Table 3.7.

Interviewers in CD	Mth 1	Mth 2	Mth 3	Mth 4
0	4	4	2	4
1	377	373	380	372
2	6	10	5	11
Total	387	387	387	387

Table 3.7: Number of Interviewers in each CD by Month

We can also consider the number of CDs from which interviewers collect data. Based on both Table 3.7 and Table 3.8 we can then determine whether interviewers are nested within CDs and also if CDs are nested within interviewers.

CDs in Interviewer	Mth 1	Mth 2	Mth 3	Mth 4
1	0	1	1	4
2-4	8	10	9	8
5-9	25	27	23	27
10-14	14	14	16	12
15+	2	1	1	2
Total	49	53	50	53

Table 3.8: Number of Collection Districts from which Interviewers Collect Data by Month

It can be seen in Table 3.7 and Table 3.8 that in the majority of cases there is only one interviewer in a single CD, there are never more than two interviewers in a single CD in any single month and that in some cases there are no interviewers in a CD in a particular month. Hence in a single month, CDs are not nested within

interviewers, as there is generally more than one interviewer in several workloads, and some interviewers (most notably the same interviewer highlighted earlier as enumerating at least part of 8 different workloads) appearing in more than one CD in a single month. This again lends credence to the assumption that this interviewer has received further training to deal with an allocation of dwellings that is very different to the other interviewers in the CURF. However in the absence of further information all we can conclude about the hierarchy is that CDs are not nested within interviewers in any single month.

In summary, in one month there are a few cases in which we have several dwellings in the same CD being allocated to two separate interviewers, while at the same time we can have two CDs being enumerated by only one interviewer. Hence the hierarchical structure is cross-classified as the person-household-CD, and person-household-interviewer structures exist side-by-side at the same level. This classification structure will be examined in more detail later in this chapter. Based on this classification structure we can appropriately structure the variance covariance matrix of response to decompose the total variance into variance components such as the interviewer effect. The following section will examine gross flows for the labour force status data item in order to examine whether there appears to be an interviewer effect in labour force status over time.

Gross Flows Analysis of Interviewers

Before applying variance decomposition models to the estimation of the interviewer effect, examination of the relevant data items may reveal some insight regarding the possible extent of the interviewer effect. We can develop an intuitive understanding as to the extent of the interviewer effect by examining the gross flows transition rates for labour force status categories by interviewer. For example, if we expect the interviewer to have a strong influence on the response collected, then for respondents who change interviewers in two consecutive months we might expect to see an increased chance that the respondent's labour force status would also change.

Table 3.9 above compares the labour force status for individual respondents over

	Employed 2nd mth		Unemployed 2nd mth		NILF 2nd mth	
Interviewer	Same	Different	Same	Different	Same	Different
Employed 1st mth	2471	4514	32	46	59	130
Unemployed 1st mth	45	64	97	161	22	67
NILF 1st mth	56	101	40	71	1146	2280

Table 3.9: Gross Flows in Labour Force Status for Consecutive Months by Change in Interviewer Status

two consecutive months of the CURF by interviewer. Of the 18,812 observations in the CURF a total of 11,402 observations could be matched for two consecutive months by the individual respondent. We can see in Table 3.9 the majority of matchable responses (65%) are collected by different interviewers in any two consecutive months. Furthermore we can see for respondents who were employed in the first month and had the same interviewer in both months, 3.5% changed their labour force status. This compares with 3.7% of respondents who were employed in the first month, changing labour force status when they were interviewed by two different interviewers. This indicates that there does not appear to be a strong effect on the response due to the interviewer. For respondents who were unemployed in the first month and had the same interviewer in both months, 40.9% changed their labour force status. Again this is similar to the 44.9% of respondents who were unemployed in the first month and were interviewed by 2 different interviewers. For respondents who were NILF in the first month the percentages are 7.7% and 7.0% respectively.

Table 3.9 provides some evidence to suggest that there may not be a strong interviewer effect associated with the labour force status data item on the CURF. However, as a summary table, this may be disguising the actual extent of the interviewer effect. For example, even though changing interviewers does not appear to lead to a respondent returning a markedly different labour force status, this could be because different interviewers influence responses in opposite directions, ultimately cancelling out any observable effect in Table 3.9.

At the aggregate level there does not appear to a strong change in the labour force

status of respondents when they change interviewers in consecutive months. This can be seen in the off-diagonal entries of Table 3.9. This information is presented in Table 3.10 for which the percentage of responses collected by different interviewers for each gross flow category is prepared. We can see that the percentage in each category is similar indicating that overall respondents do not appear to be more likely to fall into any single labour force status category if they change interviewers.

	Employed 2nd mth	Unemployed 2nd mth	NILF 2nd mth
Employed 1st mth	64.6	59.0	68.8
Unemployed 1st mth	58.7	62.4	75.3
NILF 1st mth	64.3	64.0	66.5

Table 3.10: Percentage of Total Responses in Gross Flows Categories Collected by Different Interviewers in Consecutive Months

We have seen that based on some simple exploratory analysis there does not appear to be a strong relationship between change of interviewer and month to month changes in the labour force status data item on the CURF. This simple analysis does not consider any factors that may potentially disguise the interviewer effect and hence these results may be misleading and more appropriate variance decomposition techniques can be applied to estimate the interviewer effect in practice. However, given that labour force status is the primary output from the LFS and is a derived data item compiled from approximately 80 detailed, objective questions, it appears unlikely that the influence of any single interviewer over the collected response data extends to all of the questions in the LFS. Consequently due to the complex derived nature of the labour force status data item and the objective way in which the data is compiled, we might expect any interviewer effect associated with this data item to be relatively minor. Despite this the CURF dataset presents us with an opportunity to develop methodologies and demonstrate techniques for the estimation of the interviewer effect in non-fully interpenetrated surveys.

The remainder of this chapter will consider variance decomposition techniques that can be applied to estimate the interviewer effect in practice. The following section considers the auxiliary variables that are available on the CURF and how

they can be utilized.

3.1.3 Auxiliary Variables in the CURF

In the previous sections we introduced the variables on the CURF, cleaned the data and examined the classification variables available in the dataset. There are also a number of remaining data items available on the CURF that may be considered as auxiliary variables when estimating the interviewer effect. These variables contain demographic information relating to the individual respondent such as age, gender, marital status and country of birth. These auxiliary variables can then be included as covariates during the variance decomposition procedure.

Our primary aim in this chapter is to develop and evaluated methods for the estimation of the interviewer effect associated with one or more data items on the CURF. We can do this by fitting a variance decomposition model using the data item as the response variable. In this case we are fitting a model primarily to decompose the total variance associated with the data item of interest. Hence we do not need to determine the model which has the greatest explanatory power, but only the appropriate classification structure for the decomposition of the variance. Studies have shown (e.g. Tranmer and Steel (2001); Hutchison and Healy (2001)) that ignoring a level in the classification structure during the variance decomposition results in the variation that should have been attributed to any non-included levels being distributed amongst all the remaining levels in the classification structure. Hence to estimate the interviewer effect we must incorporate the entire classification structure of the data into our variance components model, even if a simpler model may include fewer parameters and appear to describe the data more appropriately.

When estimating the interviewer effect there are a number of auxiliary variables available in the CURF that we can include as covariates. The inclusion of these variables is important as data items such as the demographic characteristics of the respondents may help to explain some of the correlation between responses that might otherwise be attributed to the interviewer. For example if we did not include age as a covariate when estimating the interviewer effect on labour force status, any

interviewer who enumerated a large number of homes containing retirees, such as a workload containing a retirement village, would most likely be interviewing a high proportion of respondents who are not in the labour force. This could lead to a high correlation in the responses collected by this interviewer and to a large interviewer effect estimate. In contrast if we control for the effect of each respondent's age on the responses collected by this interviewer we will get a more appropriate estimate of the interviewer effect.

Ultimately including the auxiliary variables in the CURF as covariates in our variance decomposition model could either reduce or increase the magnitude of our interviewer effect estimate, depending on the influence of the covariate. However we would expect this to lead to an overall reduction in the uncertainty associated with our interviewer effect estimate.

3.1.4 Summary of Exploration of CURF

The CURF dataset has been cleaned, by removing unmatchable, confidentialised missing values. We have also determined that some respondents within households have been incorrectly matched from month to month in the CURF and we have discarded these records in order to produce a consistent classification structure.

We have seen in Table 3.6 that 87% of the workloads in the CURF are enumerated by a single interviewer in any one month while 88% (see Table 3.3) of the workloads are enumerated by more than one interviewer over the four month period. Consequently we have a much higher degree of repeated measurement of workloads (or alternatively an increased degree of effective interpenetration) if we consider the CURF as a longitudinal dataset, although estimates of the interviewer effect can be produced from both the longitudinal and the single month case. It appears that interviewers who are involved in the enumeration of more than one workload in a single month may have different characteristics compared with the majority of interviewers and hence attempting to estimate the interviewer effect based on a single month will produce unreliable estimates.

The LFS has not been designed as an interpenetrating survey, however, due

to its repeated panel nature, we have a form of longitudinal interpenetration with repeated measurement of individual respondents within households over time by different interviewers. Assuming there is no change in the interviewer effect over time we can therefore use this longitudinal information to effectively increase the degree of interpenetration and produce more reliable estimates of the interviewer effect.

When estimating the interviewer effect we will initially consider simple variance components models to decompose the total variation in the data into components such as the interviewer effect. The appropriateness of more complex models incorporating covariates will also be considered in this chapter. We have already seen that the hierarchical classification structure of the data is not straightforward with responses nested within individuals, within households, within interviewers crossed with responses nested within individuals, within households, within collection districts, within workloads and a full description of this structure can be found in Figure 3.7. It is important to have a clear understanding of the hierarchical classification structure of the data so that total variation associated with our data item of interest can be decomposed appropriately during the modelling procedure. Incorporating the classification structure during the modelling procedure can also limit the number of unknown parameters to be estimated in the variance-covariance matrix and greatly speed up the estimation process. The following section considers the hierarchical structure of the data in detail before this structure is applied during the variance decomposition process.

3.2 Hierarchical Structure of the Data

This section will establish the hierarchical classification structure of the data by summarizing the exploration of the CURF in the previous sections. The classification structure will be presented using the classification diagram convention of Browne *et al.* (2001). This information will then be applied during the variance decomposition procedure to reduce the number of unknown parameters to be estimated in the

variance-covariance matrix and greatly speed up the process of estimating variance components such as the interviewer effect.

If we consider the longitudinal information as an extra level within our hierarchy it is clear that the cleaned CURF dataset has a hierarchically clustered spatial structure with responses nested within individuals, within households, within collection districts, within workloads. In contrast we have a multiple role scenario when we consider interviewers, as each individual respondent and each household may belong to multiple interviewers. To complicate matters further interviewers and workloads are cross-classified at the same level.

The basic spatial classification structure of the CURF, when considering a single month can be presented diagrammatically as in Figure 3.1 following.

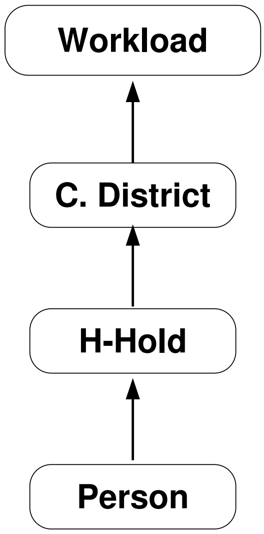

Figure 3.1: Spatial Classification Structure of CURF Data in a Single Month

Here we have depicted the hierarchical spatial classification structure of a single month in the CURF dataset using the classification diagram convention of Browne *et al.* (2001). These diagrams summarise classificatory structures by using arrows to indicate the relationship between levels in the data (depicted by enclosed polygons). A single arrow is used to indicate a hierarchically nested relationship, multiple non-parallel arrows indicate cross-classified relationships while parallel arrows indicate a multiple role/membership relationship of a lower level unit within a higher level unit. These conventions have been summarised in Figure 3.2 following.

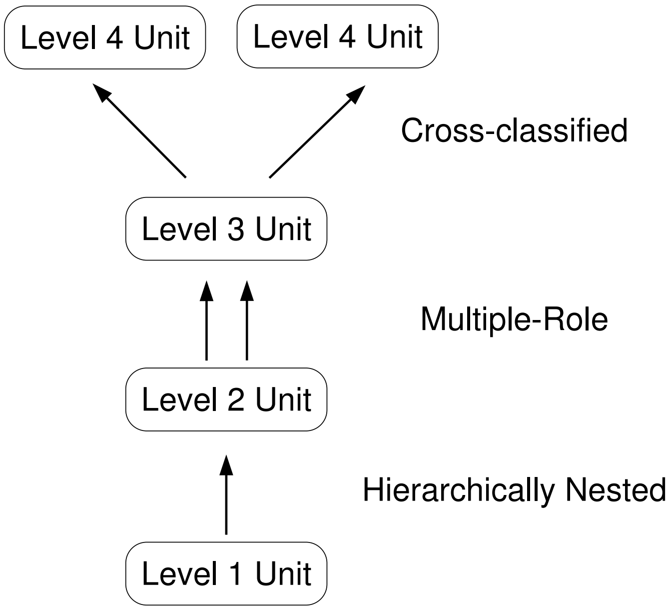

Figure 3.2: Classification Diagram Conventions

Figure 3.2 depicts all of the main classification diagrams conventions in a simple diagram. In this case we have level 1 units hierarchically nested within level 2 units which have a multiple role within level 3 units. The level 3 units are then hierarchically nested within two competing but separate classification structures at the same level (level 4). In this simple way complex classification structures of the Multiple Membership Multiple Classification (MMMC) class (introduced by Browne *et al.* (2001) as a subset of the GLMM (2.12)) can be concisely presented.

The basic interviewer classification structure in a single month of the CURF has households hierarchically belonging to interviewers, but there is no hierarchically nested relationship between interviewers and CDs or workloads.

We can see that in a single month of the CURF we have two different hierarchical classification structures in the CURF data; the interviewer structure (Figure 3.3) and the spatial structure (Figure 3.1). Both of these structures have the person and household levels in common but the interviewer level exists alongside the CD level in the spatial hierarchy. The interviewer and spatial classificatory structures can therefore be considered as cross-classified because they are separate structures that exist at the same level. This is a more complex scenario that is depicted in Figure

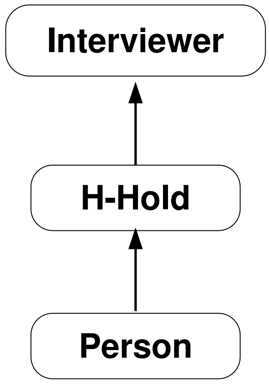

Figure 3.3: Interviewer Classification Structure of CURF Data in a Single Month

3.4 following.

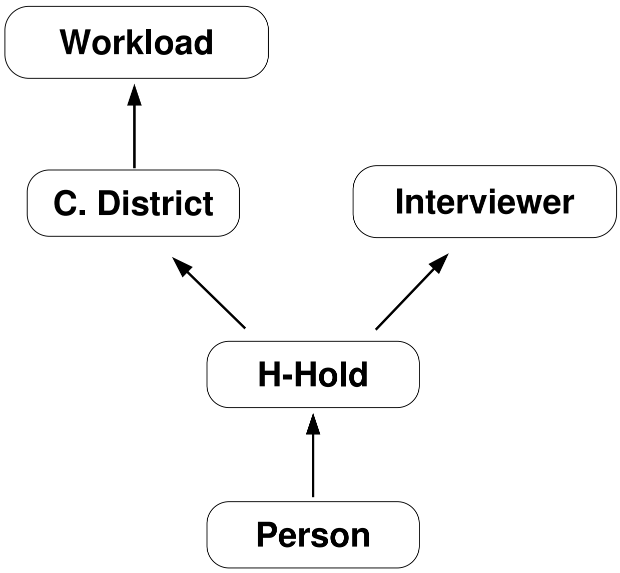

Figure 3.4: Full Classification Structure of CURF Data in a Single Month

Figure 3.4 depicts a scenario that appears to be unconfounded and standard techniques for estimating variance components at all levels of the data would appear to be applicable in this case. Unfortunately the cross-classification depicted in Figure 3.4 refers to only a very small portion (see Table 3.7) of the total interviewers and also a small portion of the total CDs and workloads and hence there is only a small amount of repeated measurement (and therefore a low degree of effective interpenetration) of either the workloads or the interviewers in this case. Furthermore the interviewers that do cross workload and CD boundaries may be extraordinary

in some way and therefore have different characteristics to the remainder of the interviewers. Hence in many cases the workload and the interviewer level are fully confounded and estimates of the interviewer effect that rely on only a single month of data will be inherently unreliable, placing undue emphasis on potentially unusual interviewers.

When we consider a single month of the CURF dataset, we are faced with a scenario in which only a small proportion of the workloads are interpenetrated. The traditional conceptualization of interpenetration (see Chapters 1 and 2) is such that if a survey is not fully interpenetrated then it is considered to be fully confounded. We can see that we do not have at least two interviewers randomly allocated to each workload in the CURF and hence the CURF is not fully interpenetrated, even if we consider the CURF as a longitudinal dataset. On the other hand we do have some interpenetrated workloads in which we get repeated measurement of the workload level spatial effect and from which we can attempt to estimate the interviewer effect. The remainder of this chapter will consider how we can estimate the interviewer effect in this *partially* interpenetrated scenario. Chapter 4 will define interpenetration comprehensively, incorporating for the first time the concept of partial interpenetration. The effect of partially interpenetrated survey designs on the precision of estimates of the interviewer effect will then be considered in Chapter 5, along with optimal design considerations for partially interpenetrated surveys.

When we consider the data by including the longitudinal information available from the repeated panel nature of the CURF we get a slightly different classification structure, from which we would expect to get more reliable estimates of the interviewer effect. If we include this longitudinal information as an extra level in our hierarchy we can consider each individual person's responses to be nested within that person. For example, a respondent who is in the survey for August, September, October and November will provide four responses to the survey.

The spatial structure over time, presented in Figure 3.5, remains purely hierarchical with each CD only ever belonging to one workload. Over time some CDs do change, however, as respondents in different CDs are rotated into the survey.

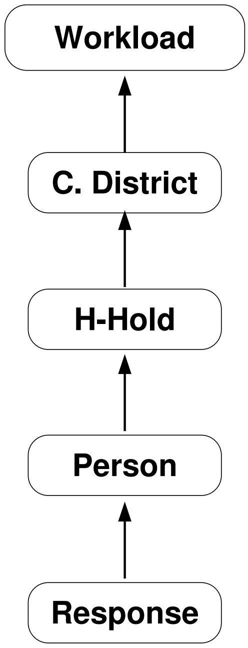

Figure 3.5: Spatial Structure of Full CURF Data

The basic interviewer structure, on the other hand, is a multiple role scenario where responses are hierarchically clustered within an individual, their household and an interviewer in any one month, but may be allocated to a number of different interviewers over time. Hence a single person or household may have a multiple role in the hierarchy, belonging to more than one interviewer. This scenario is highlighted in Figure 3.6 following.

When we consider the structure of the data, incorporating the interviewers, we then have two cross-classified structures;

1. A hierarchical spatial structure in which responses are nested within persons, within households, within collection districts, within workloads.
2. A multiple-role (or membership) structure in which responses are nested within persons, within households, but each household can belong to a number of different interviewers depending on the time period.

Note that these two structures are cross-classified at the CD level and hence also the workload level.

Figure 3.7 demonstrates that there is a complex, non-hierarchical classification

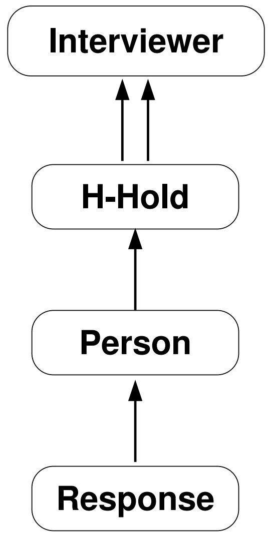

Figure 3.6: Interviewer Structure of Full CURF Data

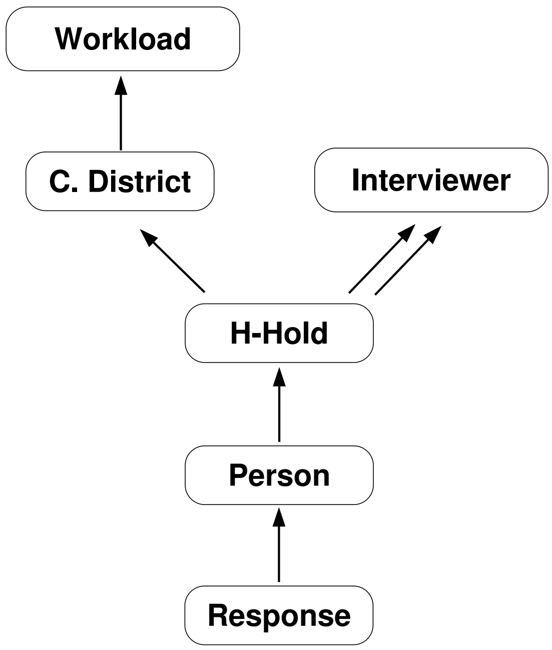

Figure 3.7: Full Structure of CURF Data

structure in the CURF. Information regarding this classification structure can now be used to appropriately decompose the variance and hence produce an estimate of the interviewer effect. Knowledge of the classification structure allows us to determine the appropriate structure of the variance-covariance matrix and greatly reduce the computational requirements of the estimation procedure.

Careful choice of the structure of the variance-covariance matrix will affect both

the number of parameters to be estimated and the estimation procedures that can be applied. In particular, for normally distributed data items, block diagonal variance-covariance matrices associated with hierarchical classification structures can be estimated using computationally efficient approximate algorithms such as IGLS. For more complex MMMC data structures variance component parameter estimates can be made using algorithms available in common software packages such as MLwiN.

3.2.1 Summary

Although the LFS has not been designed as an interpenetrated survey we can still produce estimates of the interviewer effect. If these estimates are based on a single month they will be inherently unreliable, placing undue emphasis on potentially unusual interviewers and relying on the cross-classification of a small number of units to isolate the interviewer effect. However, when an extra hierarchical level incorporating longitudinal information is included we would expect to get a higher degree of repeated measurement of both the interviewers and the workloads and hence a more reliable estimate of the interviewer effect. This is a beneficial side-effect of swapping the majority of interviewers across workloads over time, giving us a form of temporal interpenetration and repeated measurement of workloads and individuals with different interviewers.

Estimates of the interviewer effect can be made for either a single month of the CURF or for the entire CURF dataset. However the effective degree of interpenetration is much higher when we consider all four months in the CURF as a form of longitudinal interpenetration. Information regarding the classification structure of the data will then be used during the variance decomposition process to limit the number of parameters to be estimated in the variance-covariance matrix and to expedite the estimation process. The remainder of this chapter will demonstrate how we can estimate the interviewer effect in the *partially* interpenetrated, longitudinal scenario presented in Figure 3.7. Chapter 5 will then examine the general relationship between the degree of interpenetration and the precision of interviewer effect estimates.

3.3 Incorporating Longitudinal Information

In the previous section we discussed how the repeated panel information available in the CURF can be considered as an extra level in our classification structure. It is important to consider the longitudinal information in this fashion as this allows us to fully utilize temporal interpenetration to simplify the structure of the variance-covariance matrix and produce more reliable estimates of the interviewer effect compared with a single month of data. In comparison if we had fitted a time series model to the CURF, failing to specify a longitudinal level to the data, we would be trying to estimate the interviewer effect from the single month classification of Figure 3.4, ultimately producing less reliable (due to a decreased effective degree of interpenetration) estimates of the interviewer effect.

The following section extends the work of Pickery and Loosveldt (2000, 2001) who considered using repeated panel information to estimate interviewer effects in two waves of Belgian Election Studies data. Their research avoided spatial confounding by studying a single, concentrated spatial region of Belgium, which enabled them to separately apply standard hierarchical modelling techniques to estimate the interviewer effect when interviewers were not swapped over time. The following sections present an extension in which data that is not fully interpenetrated is considered and a discussion of the potential benefits that may be gained from including covariates. A simulated response variable with known random effects linked to the CURF classification structure will be used to demonstrate the potential improvement in the estimation of the interviewer effect for this case by incorporating information regarding the repeated panel nature of the LFS. A more general exploration regarding the gain from increasing the degree of interpenetration in surveys is considered in Chapter 5.

3.3.1 Relative Improvement in Interviewer Effect Estimates

The LFS has not been designed as an interpenetrated survey and so interviewer effect estimates based on a single month of data may be unreliable. This section

will simulate a normally distributed response variable with known random effects on the CURF structure to assess the potential improvement in the estimation of the interviewer effect when longitudinal information is incorporated.

A response variable with known random effects was simulated as a new column attached to the CURF dataset. The CURF level identification labels were used to additively incorporate known random effects to create this simulated variable. The following additive model was used

$$y_{ijklmn} = \mu + \omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn} + \varepsilon_{ijklmn} \tag{3.1}$$

where

- μ is a fixed effect
- The random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_{int})$, $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$ and $\varepsilon_{ijklmn} \sim N(0, \sigma^2_\varepsilon)$
- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively

The algorithm used to produce the simulated data can be found in the Appendix A. Given that the simulated response variable was normally distributed, estimates of the variance components of Model (3.1) were then made using first order Marginal Quasi Likelihood (MQL) estimation applied under the Iterative Generalized Least Squares (IGLS) algorithm in the MLwiN software package. For single months the classification structure depicted in Figure 3.4 was applied to the variance covariance matrix while the classification structure in Figure 3.7 was used in the longitudinal scenario. Thus there is an extra level in the multiple month case as there is only one measurement/response available per person in a single month. Variance component estimates based on one realization of a simulated response can be seen in Table 3.11 below.

The * in Table 3.11 refers to estimates made constraining all variance components to be non-negative. Estimated standard errors are presented in brackets.

Level	True Value	Repeated Panel	Month 1	Month 1*
Workload	0.50	0.363 (0.125)	-0.451 (0.356)	0 (0)
Interviewer	0.20	0.198 (0.041)	0.851 (0.393)	0.399 (0.110)
CD	0.75	0.709 (0.086)	0.652 (0.082)	0.652 (0.082)
Household	1.00	1.086 (0.086)	0.984 (0.074)	0.984 (0.074)
Person	2.00	1.915 (0.084)	2.018 (0.060)	2.018 (0.060)
Response	4.00	4.031 (0.053)	-	-
Fixed Effect	10.0	10.0 (0.1)	10.0 (0.1)	10.0 (0.1)

Table 3.11: Comparison of Single Month and Repeated Panel Variance Component Estimates for One Realization of Simulated Response

Table 3.11 demonstrates that in this example we get poor estimates of the interviewer effect in the weakly interpenetrated case of a single month. However repeated panel information can be incorporated to greatly increase the degree of interpenetration and improve estimates of both the interviewer effect and the workload level variance component.

The estimates presented in Table 3.11 have been simulated based on the assumption interviewer effects derive from a normal distribution. We have already speculated that interviewers crossing workloads in a single month may be different to the majority of interviewers and hence estimates based on a single month of the CURF are likely to perform even less well in relation to the repeated panel scenario than has been suggested in Table 3.11. Based on the simulated results we can see in Table 3.11 that the variance component estimates are more stable in the repeated panel scenario. Although variance component estimates are similar at the person, household and CD levels, the interviewer and workload level variance component estimates are much closer to the simulated value in the repeated panel scenario than for a single month. This is due to the low degree of interpenetration in single months of the data for which the interviewer and workload level are almost confounded. This leads to unreliable interviewer and workload level variance component estimates in the single month scenario. In comparison the interviewer and workload level variance component estimates are much more appropriate in the repeated panel case. It can also be seen that the longitudinal estimates generally

have a lower estimated standard error, although this would be expected as they are based on approximately 4 times the amount of data as the single month estimates.

Although results from this simulation are only indicative as there are no continuous data items available in the CURF, Table 3.11 clearly demonstrates that in this example there is a large potential gain for estimating the interviewer effect from including the longitudinal information available in repeated panel surveys. We would expect to see a similar relative gain when we consider the real data. This raises the question as to the effect of partial interpenetration on interviewer effect estimates and whether partially interpenetrated designs may be more cost effective than fully interpenetrated designs for the estimation of the interviewer effect. Chapter 5 will explore this issue in more detail while the remainder of this chapter will demonstrate how the interviewer effect can be estimated on the CURF dataset.

3.4 Variance Decomposition

This section will estimate the interviewer effect on labour force status in the CURF dataset by using longitudinal interpenetration to increase the effective degree of interpenetration during the variance decomposition procedure. The structure of the CURF is complex and highly non-hierarchical. Consequently we will consider MMMC variance decomposition models to estimate the interviewer effect.

Labour force status can be considered as an unordered multinomial response variable with three categories; employed, unemployed and Not In the Labour Force (NILF). We will initially consider a binary employment status variable created by a transformation of labour force status. The binary employment status variable was created by summing the unemployed and NILF categories producing a variable with only two categories; employed and not employed. This binary employment status variable can then be modelled as a binomial response. The following section will discuss a number of issues associated with estimating variance components with binary and multinomial response variables.

3.4.1 Estimation in Logistic Variance Decomposition Models

This section presents a brief review of variance component estimation with binary response variables and discusses the estimation techniques that will be applied to estimate the interviewer effect associated with labour force status on the CURF dataset.

Although approximate techniques for the estimation of random effects perform well with normally distributed response variables it has been shown that first order MQL estimation is biased for binary response variables (e.g. Gilmore *et al.* (1985); Breslow and Clayton (1993); Rodriguez and Goldman (1995) and Rodriguez and Goldman (2001)). Improvements to approximate techniques such as second order Penalised Quasi-Likelihood (PQL) estimation (see Goldstein and Rasbash (1996)), have reduced but not eradicated this bias (see Breslow (2003) for a review). Rodriguez and Goldman (2001) showed that both first and second order MQL and PQL estimates will generally exhibit greater bias if either

- The random effects are large, or
- There are generally small numbers of level 1 units within level 2 units

It is likely we will be faced with both of these scenarios when considering the CURF. In particular there are only an average of approximately two respondents within each household.

More recent approximate estimation techniques such as the h-likelihood estimators associated with the Hierarchical Generalized Linear Models (HGLMs) of Lee and Nelder (1996) are also biased for non-linear response variables. Furthermore Kuk and Cheng (1999) and Waddington and Thompson (2004) found that h-likelihood estimators are unsatisfactory for some parameters when applied to binary response variables.

In comparison both Rodriguez and Goldman (2001) and Breslow (2003) have shown that exact methods of estimation such as adaptive Gaussian quadrature, stochastic integration and Bayesian estimation can be applied to produce unbiased

random effect estimates for binary response variables. Of these options Breslow (2003, p ii) suggests that

'*MCMC is likely to be the method of choice for the most complex problems that involve high dimensional integrals.*'

and this approach will be considered in this chapter.

Given several techniques that can be applied to estimate random effects of unknown extent on a binary response variable it may be difficult to evaluate the appropriateness of any individual estimate. Approximate techniques may suffer from a bias that can ultimately be corrected by asymptotically unbiased bootstrapping techniques, while exact techniques may return inappropriate results if applied without caution, such as if a MCMC chain is halted prematurely.

The question then is how to evaluate which estimation procedure produces the most appropriate estimate. Rodriguez and Goldman (2001) suggest comparing estimates from a number of different methods to increase support for the final results. In particular if there is a downward bias produced by a first order approximate procedure, the second order estimates should be less biased and hence closer to any estimates produced by an exact methodology. Bootstrapping of either the first or second order estimates can then be used to further support the estimates produced using the exact procedure.

Demonstrating Expected Bias Through Simulation

Rodriguez and Goldman (2001) highlighted the bias that can occur when estimating variance components for binary response variables and suggested this bias is related to both the magnitude of the variance components and design factors such as the number of level one units within level two units. Consequently we can demonstrate for any survey design, whether we might expect approximate variance component estimates made under a variance decomposition model, to be biased. This can be done by simulating a binary response variable comprised of random effects of known magnitude according to the relevant variance decomposition model and the classification structure implied by a given design. If this is done after the data has

been collected, the simulated effect size can be set as the variance component estimates derived from the real data. Variance component estimates based on both approximate (such as MQL and PQL) techniques and exact techniques can then be compared to examine whether a given survey design and variance component magnitude indicates a potential bias in estimates made under approximate techniques.

3.4.2 Practical Application of MCMC to Estimation of Variance Components

MCMC estimation of variance component parameters in GLMMs is a technique that must be applied with care or it may result in misleading estimates. Consequently in this section we will describe a number of systematic steps that will be applied in the MCMC estimation of the interviewer effect on non-linear data items in this chapter. These steps will be described in relation to the most common areas of inappropriate application of MCMC estimation techniques.

Appropriate Prior Specification

The specification of an appropriate prior distribution is one of the most contentious areas of MCMC estimation of variance component parameters. For example, apart from computational intensity, Noh and Lee (2004, p 2) highlight the possibility that simulation-based methods

'*...could result in wrong estimates, which may not be detected*'

However this statement relies on the results of Hobert and Casella (1996) who showed that with improper priors it is possible for commonly used convergence diagnostics to be misleading. Hobert and Casella (1996) conclude that careful specification of priors and appropriate use of diagnostics can be used to determine the appropriateness of a specified prior distribution.

In this chapter we will initially apply the diffuse inverse Wishart prior distribution for the variance parameters as developed in Browne and Draper (2000). These prior distributions for the estimation of variance component parameters have been

implemented in MLwiN 2.0. Alternative prior distributions will only be considered if post-estimation examination of the posterior indicates that the prior is inappropriate.

Appropriate Choice of Starting Values for MCMC Chain

We discussed in Chapter 2 how the choice of initial values may influence the convergence speed of the MCMC chain. One of the indicators that a MCMC chain may have achieved convergence is that a number of chains initialized from different starting values all converge to the same estimates. In this chapter a minimum of three sets of starting values are applied in the estimation of parameters for each separate model, as listed below

1. The initial values for the parameter estimates are set to be the parameter estimates under second order MQL estimation. Note that for cross-classified models, model construction is different for MCMC and MQL/PQL estimation and hence MQL/PQL starting values are generally entered manually.

2. The initial values for the parameter estimates are set to be the parameter estimates from the MCMC chain initialized on the MQL estimates. Note that this makes sense as for strongly biased MQL estimates the adaptive Metropolis-Hastings settings in MLwiN may not be appropriate based on a single run.

3. The initial values for the parameter estimates are set to be the parameter estimates from the MCMC chain initialized on the first MCMC estimate plus the difference between the parameter estimates from the MCMC chain initialized on the first MCMC estimates and the parameter estimates initialized on the MQL estimates. If there is any bias in the approximate MQL estimates it would be expected that this set of initial values will have a positive bias and hence we would hope to see convergence from above in this case.

The stability of the final estimates can then also be confirmed by initializing the MCMC chain using the final estimates as starting values. With complex variance

decomposition structures the interaction of the different variance components in the model may also inhibit convergence so it is generally also sensible to implement several sets of starting values in which one of the variance components is set to a random starting value.

Determining Convergence: When to Stop a MCMC Chain

In practice determining when to halt a MCMC chain will generally be a pragmatic decision based on the computing time required and the desired level of accuracy of the estimates. In general MCMC chains in which the starting value is close to the final estimate will generally converge faster but other techniques such as thinning, hierarchical centring, reparameterization and altering the proposal distribution can be used to increase convergence speed (see for example Browne (2004)).

Post-run diagnostics can then be used to determine if the MCMC chain has converged satifactorily. Examination of the trajectories for each parameter estimate will indicate whether the estimate has stabilized or if the MCMC chain needs to be run for further iterations to achieve convergence. Similarly examination of the Auto-Correlation Function (ACF) and Partial Autocorrelation Function (PACF) plots will indicate high levels of autocorrelation in the MCMC chain requiring a longer run to achieve estimates with a desired level of accuracy. Diagnostic tools which consider the autocorrelation levels of the MCMC chain can also be used to indicate the expected number of iterations required to estimate the parameters to a specified degree of accuracy.

Summary

The steps outlined above, combined with post-run diagnostic checks do not *guarantee* the effectiveness of MCMC methods for the estimation of variance component parameter. However, these steps do allow us to avoid many of the common pitfalls through which MCMC estimation can give misleading results and increase the support for the final estimates. With non-linear models, for which the approximate estimation techniques are biased, further support for the MCMC estimates can be

gathered through post-estimation simulation of a response variable with known random effects based on the MCMC estimates.

Although it is ideal to compare several different estimation procedures, computational time is an important consideration when considering large datasets. Exact estimation methods are generally more computationally intensive than approximate techniques and this is also an issue when bootstrapping. The following section will explore the binary employment status case, before we determine an appropriate estimation procedure for the more complex multinomial response scenario.

3.4.3 Variance Decomposition of Binary Employment Status

In this section we will produce some preliminary estimates using binary response variance component models to get an indication as to the extent of the interviewer effect in the CURF dataset. Later sections will consider the full labour force status variable using an expanded multinomial multilevel analysis.

Given binary employment status, y, we can model the probability that a person's responses will be in either state 0 (employed) or state 1 (not employed) with the following multilevel logistic regression model form of the GLMM

$$\Pr(y_{ijlkmn} = 1|\pi_{jklmn}) = \frac{\exp(\pi_{jklmn})}{1 + \exp(\pi_{jklmn})} \tag{3.2}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the CURF

$$\pi_{jklmn} = \mu + \omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn}$$

and

- μ is a fixed effect
- The random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_{int})$ and $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$

- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively. Note that these indices do not, by themselves, detail the hierarchical structure of the data
- Information regarding the hierarchical classification structure of the data is presented in Figure 3.7

Information regarding the temporally interpenetrated classification structure (Figure 3.7) of the CURF was utilized to determine the appropriate estimation procedure for Model (3.2) and how it could be applied in practice. Estimates of the variance components were then made using a number of different estimation procedures available in MLwiN 2.0. Results from these procedures can be seen in Table 3.12 following.

Level	MQL (1)	MQL(2)*	MQL (Boot)	MCMC
Workload	0.05 (0.03)	0.53 (0.11)	0.59	3.12 (2.27)
Interviewer	0 (0)	0 (0)	0.002	0.10 (0.09)
CD	0.20 (0.04)	0.89 (0.13)	2.18	16.02 (4.04)
Household	1.04 (0.07)	4.28 (0.22)	11.3	84.18 (10.11)
Person	1.04 (0.06)	7.65 (0.19)	11.3	107.94 (10.99)
Fixed	0.44 (0.05)	1.48 (0.14)	2.81	3.81 (0.46)

Table 3.12: Variance Component Estimates for Binary Employment Status: Repeated Panel Data

In Table 3.12

- MQL (1) stands for first order marginal quasi-likelihood estimation. Note that variance component estimates have been constrained to be non-negative, hence an estimate of 0 (0) indicates that the initial estimate was negative
- MQL (2) for second order marginal quasi-likelihood estimation. The * here is used to indicate that there were numerical errors and that the MQL(2) procedure failed to converge. The estimates reported are indicative values produced on the final iteration before numerical problems forced the MQL(2) procedure to be halted

- MQL (Boot) for bootstrapped MQL estimates (see Kuk (1995) for more information). Note that standard error estimates are currently not available for this procedure; and

- MCMC for Bayesian estimates produced using MLwiN's adaptive Metropolis-Hastings MCMC algorithm. Note that the standard errors presented in Table 3.12 are not standard errors in the traditional sense, but rather a Bayesian equivalence based upon empirical sampling from the posterior distribution, which will approximate the standard error when the posterior is normally distributed

We can see in Table 3.12 that we get different variance component estimates depending on the estimation procedure applied. Moreover we can see that the first order MQL estimates are the lowest, followed by the second order MQL and the bootstrapped MQL estimates. This is the pathology Rodriguez and Goldman (2001) identified as indicating that the underlying variance components are large, leading to a downwards bias in the approximate estimates. Rodriguez and Goldman (2001) also found that estimates made using PQL exhibited the same pattern of bias. However these estimates have not been included after fatal numerical problems in estimation. In this case the bias corrected estimates made using the bootstrap should eventually converge on the exact estimates made using the MCMC sampler. However this is a very computationally intensive process (much more so than the MCMC sampler) and the bootstrapping procedure was halted after the downward bias had been identified but before convergence could be achieved.

The MCMC estimates were fitted from a number of different starting values which consistently converged to estimates of similar magnitude. Thus there is considerable evidence to suggest that the underlying variance components of the binary employment status variable are large and that the approximate estimates exhibit a strong downward bias. This is as would have been expected given generally low numbers of respondents within households. Rodriguez and Goldman (2001) highlight this as a cause of bias in logistic multilevel models. Given the evidence of bias

in the approximate estimates in Table 3.12 the strongest support would be given to the estimates produced using the MCMC sampler in MLwiN 2.0 and these results will now be examined in more detail.

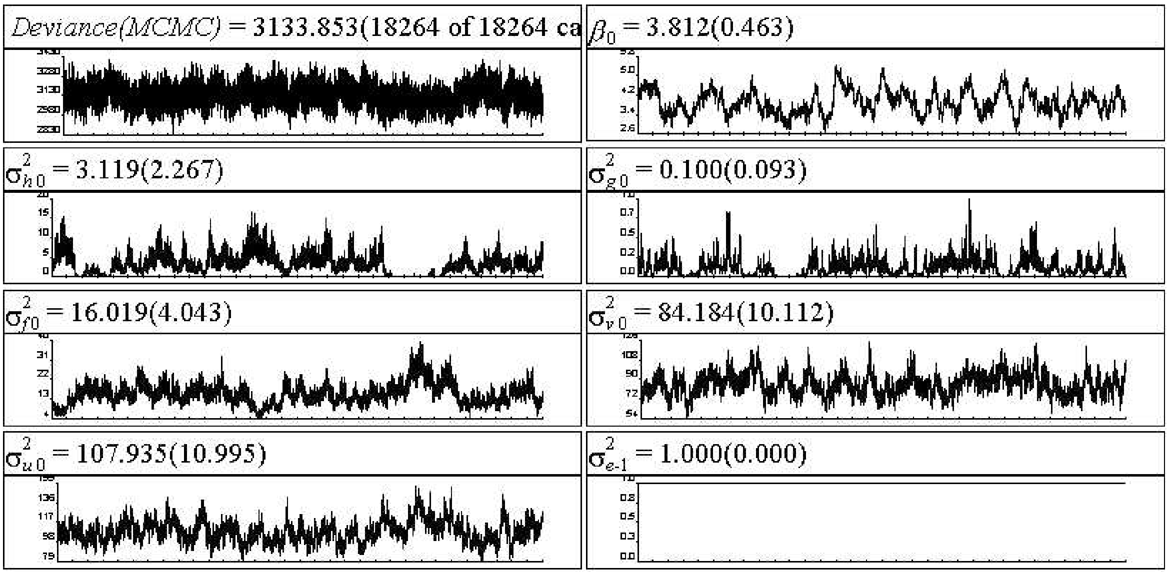

Figure 3.8: MCMC Trajectory Plots for Variance Components of Binary Employment Status: Repeated Panel Data

Figure 3.8 shows the estimates for each of the parameters in our binary response variance components model (3.2) along with a Bayesian equivalence (see Spiegelhalter *et al.* (2002)) to the deviance statistic of McCullagh and Nelder (1989) for each of the last 32,000 iterations of the MCMC sampler in MLwiN 2.0. In general these Gibbs sampling traces do not look particularly healthy for any of the parameters as there appears to be some auto-correlation between succeeding iterations in the MCMC chain. As Browne (2002, p 25) points out sampling traces

'*... when considered as a time series these traces should resemble 'white noise'.*'

and we can see in Figure 3.8 that none of the traces appear to resemble white noise. This suggests some autocorrelation between successive estimates in the MCMC chain. High levels of auto-correlation indicate that the MCMC chain is mixing badly and will have to be run for a greater number of iterations to produce estimates of a required accuracy. A number of diagnostics tools can be used to examine the effect of autocorrelation on the estimate of the interviewer effect and the standard MLwiN

2.0 diagnostics have been included in Figure 3.9 following.

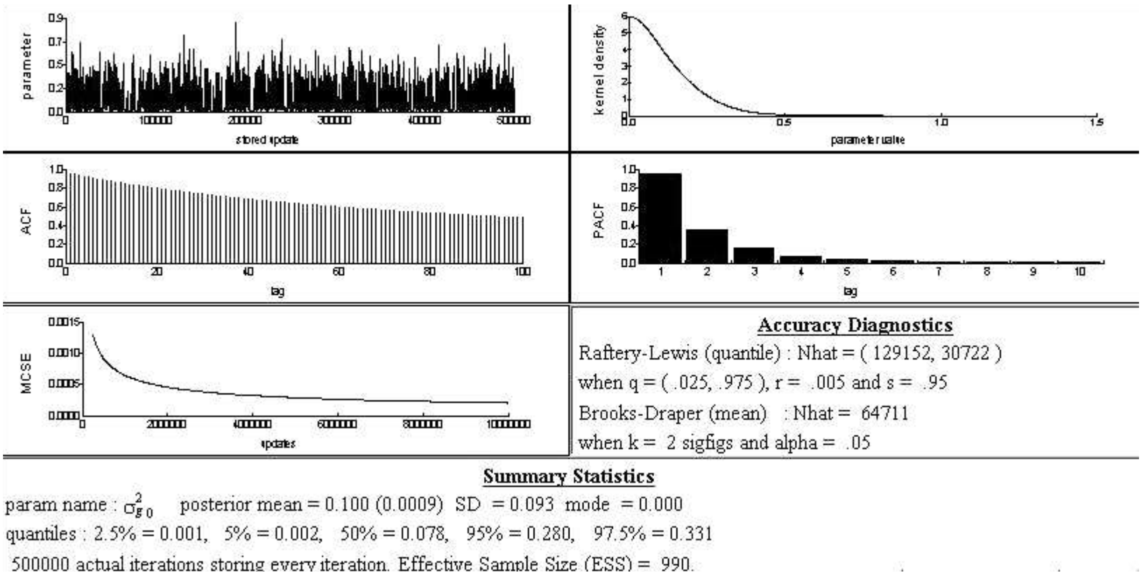

Figure 3.9: MCMC Diagnostic Plots for Interviewer Effect of Binary Employment Status: Repeated Panel Data

In Figure 3.9 we can see from the parameter trace (upper left corner), AutoCorrelation Function (ACF) and Parital AutoCorrelation Function (PACF) plots that there is a high degree of autocorrelation between succeeding iterations in the MCMC chain. Poor mixing is a violation of a property of Markov chains and indicates an effective reduction in the number of independent draws made using the MCMC sampler. By limiting the effective sample size, high levels of autocorrelation reduce the convergence speed of the MCMC sampler, although this does not alter the properties of the estimated posterior distribution once convergence has been achieved.

Based upon this autocorrelation, the Raferty-Lewis and Brooks-Draper diagnostics can then be used to indicate how many iterations of the MCMC sampler are needed to produce estimates of a required accuracy. Here the Brooks-Draper diagnostic indicates that we would need to run the MCMC algorithm for only 64,711 iterations to be able to quote the interviewer effect estimate to two significant figures. This is a far shorter chain than the 500,000 iterations presented in Figure 3.9 and consequently we can see that the autocorrelation in the MCMC chain is

not affecting our estimates too badly. Similarly the Raferty-Lewis diagnostic indicates that the chain should be run for approximately 130,000 iterations in order to estimate quantiles to the same degree of accuracy.

We can see in Figure 3.9 that we have run the MCMC chain for enough iterations to quote our interviewer effect estimate to two significant figures. However if the autocorrelation of the MCMC chain was too high we could produce more accurate estimates of the interviewer effect by

- Run the MCMC chain for longer; i.e. until we have sampled enough effectively independent draws from the posterior distribution to estimate the interviewer effect to the desired degree of accuracy. The Brooks-Draper diagnostic can be used to give an indication as to the required length of run, based on the observed autocorrelation of the MCMC chain, to produce estimates to two significant figures
- Transform either the model or the proposal distribution in order to improve mixing and convergence speed
- Fit the model using an alternative MCMC sampler. For example Browne (1998) highlights that in extreme cases Gibbs sampling will generally result in a MCMC chain with less correlation than the standard adaptive Metropolis-Hastings sampler of MLwiN 2.0. Furthermore Browne and Draper (2003) suggest that the adaptive rejection sampler of Gilks and Wild (1992) implemented in WinBUGS (Spiegelhalter *et al.* (1994)) may also be preferred in some circumstances. In general Browne and Draper (2000) showed that the adaptive Metropolis-Hastings algorithm is the most appropriate sampler for logistic multilevel models
- Thin the stored chain. This technique stores only every kth iteration of the MCMC chain. It offers only marginal increases in convergence speed but can be used to greatly reduce computational storage requirements

However given Bayesian standard error estimates for the interviewer effect, we only need to adopt these techniques if we desire a specified level of accuracy. Based on the

assumption (which can be assessed through evaluation of the diagnostics) that the posterior is normally distributed, quoting the Bayesian standard error estimates will discourage inappropriate usage of the interviewer effect estimates and indicate the level of confidence that can be associated with estimates derived from a shortened MCMC chain. Note that if the posterior is not normally distributed quantiles should be quoted instead of the Bayesian standard error.

Similar diagnostics tools can be used to examine the other variance components in Model (3.2). It was found that the highest autocorrelation was in succeeding iterations of the interviewer effect and hence the MCMC convergence speed was determined by this parameter. Consequently diagnostics for the remaining variance components in Model (3.2) have not been presented in this chapter.

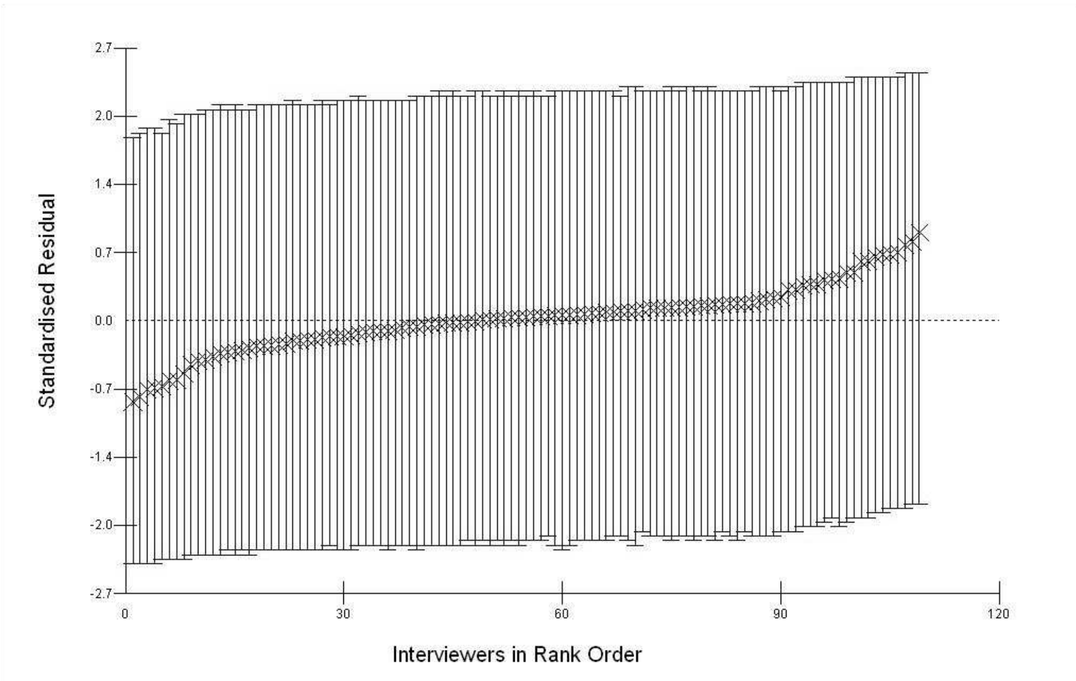

Figure 3.10: Ranked Interviewer Residuals Plot

We can then examine the interviewer level residuals, which can be saved in MLwiN, to further assess the model fit and identify outliers, which in this case would relate to unusual (or in some way exceptional) interviewers. In Figure 3.10 we can see that none of the interviewers appear to be significantly different from the others. This is due to the small size of the CURF dataset, which is a sample of 50

workloads drawn over 4 consecutive months from a much larger survey. In order to distinguish between set of responses collected by individual interviewers we would need to include further data. Note that extra data is available on the complete LFS dataset from which the CURF was derived.

In this case when we refer to *exceptional* interviewers we are considering a concept similar to that of Pickery and Loosveldt (2004, p 77) in which they consider exceptional interviewers to be those that

'*...register unusual response patterns compared to other interviewers.*'

It is important to be able to identify exceptional interviewers as this information can be used as part of a survey monitoring and maintenance program and to target interviewer training. This could potentially occur in real time, such as over successive waves of a repeated panel survey, allowing the survey methodologist to adjust for interviewing problems as they occur.

Identifying Exceptional Interviewers

We had previously seen that there was one interviewer in the CURF dataset that appeared to deal with a markedly different allocation of respondents than the other interviewers. We can now identify that interviewer by their ranking in Figure 3.10 to determine whether the responses collected by this interviewer appear to be exceptional in some way with respect to the interviewer level residuals. Although we have seen that we do not have enough data to distinguish between interviewers based on the interviewer level residuals, we can still explore how the characteristics of specific interviewers relate to the interviewer effect.

Interviewer id	Rank	Persons	Wklds	CDs	Hhs
14455	109	63	1	4	34
25516	1	133	1	7	62
37720	48	53	8	9	24
30095	85	181	3	16	92
Mean over all interviewers		114.9	1.2	8.6	56.5
Median over all interviewers		107	1	8	54

Table 3.13: Characteristics of Selected Interviewers

In Table 3.13 we can see that the interviewer level residual for the interviewer who travels to 8 different workloads is ranked 48th out of the 109 interviewers. This suggests that although the way in which respondents have been allocated to this interviewer may be unusual, the data this interviewer collects from respondents does not appear to be unusual compared with the data collected by other interviewers. Consequently we could not say that this interviewer appears to have a significantly different effect on the responses that he or she collects compared with the other interviewers.

In a similar way, other individual interviewers can be highlighted and we can examine how the characteristics of the interviewer or the way in which respondents are allocated to the interviewer relates to the interviewer effect. This could be done, for example, by examining the relationship between the magnitude of the interviewer level residual and the interviewer and workload characteristics available in the dataset. The interviewer residual ranking across a number of different questions, or data items in a single survey can also be compared in a similar way. Note that the variance decomposition model applied in this section considers the interviewer effect as a normally distributed random effect at the interviewer level. Consequently any remaining structure in the interviewer level residuals, for example such as if the interviewer effect for particular interviewers is related to the age of the interviewer, would suggest that this information should be included as an interviewer level covariate in our model. Although information on interviewer characteristics is not available in the CURF, this would be a simple extension allowing us to consider how the characteristics of both the interviewer and the survey design structure influence the data collected from respondents.

Comparing variance decomposition models in order to determine the most appropriate interviewer and individual level covariates to include in a variance decomposition model is still a somewhat controversial process under the GLMM. Consequently it is not immediately clear how to determine whether a particular covariate should be included in the model. This issue will be examined in greater detail when the gain from including covariates is examined later in this chapter.

Summary

We have demonstrated in this section how to estimate the interviewer effect for a non-fully interpenetrated repeated panel survey, using the longitudinal information as an extra level in the variance decomposition model to increase the effective degree of interpenetration in the dataset. Although the interviewer effect estimate was non-significant, in this case, this is the first empirical demonstration of how to estimate the interviewer effect in a partially interpenetrated survey. The impact of partial interpenetration on the estimation of the interviewer effect will be considered in more detail in Chapters 4 and 5.

We have seen that when approximate techniques, such as MQL, are utilized for the estimation of the variance components of a binary employment status variable there appears to be a strong downward bias in these estimates. In comparison there is strong support for estimates of the interviewer effect made using exact MCMC techniques. We initially decomposed the variance in order to estimate the component of the total variation that can be attributed to the presence and characteristics of the interviewer. Following the work of Pickery and Loosveldt (2004) we then demonstrated how examination of the interviewer level residuals allows the identification of unusual or exceptional interviewers. Although there is not enough data in the CURF to distinguish individual interviewers, under the simple variance decomposition model (3.2), we have seen that highlighting exceptional interviewers, data items and questions will lead to improvements in data quality. In particular, for repeated panel surveys, monitoring the interviewer level residuals could potentially occur as a regular part of the survey process, allowing survey designers to make rapid changes to interviewer training as potential problem areas are highlighted.

The following sections will further explore the appropriateness of the MCMC estimates by evaluating the performance of a number of estimation techniques on a binary response variable with known random effects simulated on the classification structure of the CURF dataset.

3.4.4 Estimated Bias on Simulated Dataset

Although the MCMC procedure appears to have produced the most appropriate variance component estimates, we can further validate these results by comparing the performance of the different estimation techniques on a simulated response with known random effects. Here we apply the MCMC variance component estimates in Table 3.12 to the CURF data structure to produce a simulated binary employment variable. An algorithm to simulate a binomial response with known random effect sizes on the CURF structure can be found in Appendix A.

Level	Simulated Effect	MQL (1)	MQL (Boot)	MCMC
Workload	10	0.09 (0.03)	0.81	10.1 (3.9)
Interviewer	0.01	0 (0)	0.027	0.018 (0.025)
CD	20	0.25 (0.04)	2.15	21.4 (4.1)
Household	85	0.98 (0.07)	8.64	89.6 (11.3)
Person	115	1.14 (0.07)	11.1	117.7 (12.6)
Mean	5	0.57 (0.06)	4.52	5.49 (0.64)

Table 3.14: Variance Component Estimates for One Realization of a Simulated Repeated Panel Binary Response with Known Random Effects

Note that in Table 3.14 the same replicate set size, maximum number of iterations per replicate and maximum number of sets options were chosen for the bootstrap options as in Table 3.12.

We can see in Table 3.14 that based on simulated random effects of a known magnitude the approximate MQL estimates have a large downward bias. In comparison the MCMC random effect estimates are all within one standard error of the simulated value. This simulation presents similar results to those found in Table 3.12 confirming the expected downward bias from approximate estimation techniques on the CURF structure with binary response and demonstrating the appropriateness of MCMC estimation techniques in this case.

We have shown for the case of a simulated binary response variable with known random effects of similar magnitude to the MCMC estimates presented in Table 3.12 that MCMC estimation gives variance components estimates within one standard

deviation of the simulated value. In this example the interviewer effect was small in comparison to the spatial effect and the following section will assess the ability of the MCMC techniques to estimate the interviewer effect when the interviewer effect is large relative to the spatial effects.

Estimability of Large Interviewer Effects

Although we have demonstrated the applicability of MCMC techniques for estimating variance components on the CURF, it is possible that we may not have been able to fully separate the interviewer effect from the spatial effects leading to a potential underestimate of the interviewer effect. This section will consider a simulated binomial response variable on the CURF structure with a large interviewer effect in relation to the spatial effect terms, in order to determine the appropriateness of the MCMC models when the interviewer effect is significant. An issue with MCMC estimation is the appropriate choice of initial values as this can affect the convergence speed of the MCMC chain and may lead to inappropriate estimates for a MCMC chain of a given length.

		Starting Values for MCMC Chain		
Level	Simulated Effect	MQL(1)	MQL(1)*	Simulated Values
Workload	5	0.3 (0.5)	2.7 (3.1)	2.5 (3.1)
Interviewer	10	2.4 (1.1)	11.5 (2.5)	11.4 (2.5)
CD	20	3.7 (1.5)	20.2 (4.0)	20.3 (4.1)
Household	85	21.3 (6.6)	81.0 (9.5)	81.0 (9.6)
Person	115	11.0 (6.6)	105.0 (10.9)	104.7 (10.7)
Fixed	5	2.0 (0.5)	4.6 (0.6)	4.6 (0.6)

Table 3.15: MCMC Variance Component Estimates for One Realization of a Simulated Repeated Panel Binary Response with Interviewer Effect of Increased Magnitude

Note in Table 3.15 the * refers to a re-initialization of the MCMC chain based on the initial MQL(1) estimates. The MQL(1)* estimates therefore apply more appropriate initial values for the MCMC chain. We can see in Table 3.15 that we get similar variance component estimates no matter whether we use the simulated effect size as the initial value or the MQL(1) estimates. Convergence to the same

posterior (and the same estimates) from a number of different starting values lends support to the MCMC estimates as it suggests that the MCMC chain has converged appropriately and that we have not instead reached a local mode while we wait for the MCMC chain to move on. We can also see that MCMC estimates initialized from the MQL(1)* and the simulated values are all within one Bayesian standard error of the the actual simulated values. Thus we can see that, in this example, careful application of MCMC estimation techniques produce appropriate estimates of the interviewer effect even when the interviewer effect is significant.

Note also that if we did not know the *true* magnitude of the variance components we can still identify inappropriate MCMC estimates by initializing the MCMC chain from a number of different starting values. An examination of the MCMC trajectories will also indicate a lack of convergence as a continuing trend in the trajectory will indicate the chain has yet to approach a stable estimate.

Summary

In this section we demonstrated how the longitudinal information available in the CURF dataset can be utilized in the estimation of the interviewer effect. This is the first empirical study we are aware of that appropriately estimates the interviewer effect for a binary response variable drawn from a large regularly run partially interpenetrated repeated panel household survey.

Although we can use repeated panel information to produce improved estimates of the interviewer effect, a number of issues must be considered when estimating variance components on binary response variables. Estimates produced using approximate techniques have a downward bias, while exact techniques are computationally intensive and may converge slowly. Comparisons of estimates made using a number of different techniques indicate that the interviewer effect on binary employment status in the CURF data is small in relation to the spatial effects. It would be expected that further extensions to these estimation techniques, such as in the multinomial labour force status case, will produce similar results but also be affected by the same estimation issues. The next section will extend this analysis

to consider full labour force status as a response variable. Given that we would again expect a large downward bias when applying approximate estimation, exact estimation techniques will be applied exclusively in this section.

3.4.5 Variance Decomposition of Full Labour Force Status

This section will extend the binary response variable analysis of the previous sections to consider the full labour force status response variable. The full labour force status variable, on the CURF, has three response categories, employed, unemployed and Not In the Labour Force (NILF). We can consider labour force status as an unordered multinomial response variable and estimate the interviewer effect associated with this data item by using a multinomial extension to the multilevel logistic variance decomposition model (3.2) introduced in the previous section.

In the logistic model we considered the probability that the response was not employed compared to that of being employed. In the multinomial case we choose a baseline category and model the probability that the response is any other state, i, in relation to this baseline. Given our multinomial labour force status, y, and setting employment as the baseline category (state 0) we can model the probability, $\mathrm{P^{(i)}}$, that a person's responses will be in one of t non-baseline states i (in this case we have $t = 2$ alternate states: unemployed when $i = 1$ and NILF when $i = 2$) with the following multilevel multinomial regression model (3.3)

$$\mathrm{P^{(i)}} = \mathrm{Pr(y_{qjlkmn} = i}|\pi^{\mathrm{(i)}}_{\mathrm{jklmn}}) = \frac{\exp(\pi^{\mathrm{(i)}}_{\mathrm{jklmn}})}{1 + \sum_{\mathrm{i=1}}^{\mathrm{t}} \exp(\pi^{\mathrm{(i)}}_{\mathrm{jklmn}})} \tag{3.3}$$

where $\pi^{(i)}$, can be decomposed into a number of variance components corresponding to the classification levels in the CURF

$$\pi^{(i)}_{jklmn} = \mu^{(i)} + \omega^{(i)}_n + \nu^{(i)}_{mn} + \varphi^{(i)}_{lmn} + \phi^{(i)}_{klmn} + \theta^{(i)}_{jlkmn}$$

and

- The random effects are all independent and normally distributed, i.e. $\omega^{(i)}_n \sim N(0, (\sigma^{(i)}_\omega)^2)$, $\nu^{(i)}_{mn} \sim N(0, (\sigma^{(i)}_\nu)^2)$, $\varphi^{(i)}_{lmn} \sim N(0, (\sigma^{(i)}_\varphi)^2)$, $\phi^{(i)}_{klmn} \sim N(0, (\sigma^{(i)}_{int})^2)$

and $\theta^{(i)}_{jklmn} \sim N(0, (\sigma^{(i)}_{\theta})^2)$. Note that in this case each of the variance components are not assumed to be independent for each state, i or constrained to be equal across the states, though these assumptions may be applied to reduce the complexity of the model

- q, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively
- The classification structure of the data can be found in Figure 3.7

Then as we have specified all possible states, the probability of being in the baseline state can be simply obtained by subtraction.

$$\mathrm{P}^{(0)} = 1 - \sum\nolimits_{\mathrm{i}=1}^{\mathrm{t}} \exp(\pi^{(\mathrm{i})}_{\mathrm{jklmn}})$$

Estimates of the variance components were then made using the MCMC engine available in MLwiN 2.0. Results can be seen in Table 3.16 following.

Level	Unemployment	NILF	Covariance
Workload	0.217 (0.559)	1.066 (1.387)	0.278 (0.816)
Interviewer	0.156 (0.127)	0.133 (0.102)	0.036 (0.086)
CD	9.364 (2.431)	21.01 (4.569)	13.00 (3.116)
Household	22.33 (10.39)	98.06 (25.73)	45.90 (17.69)
Person	57.12 (11.83)	132.5 (15.92)	79.17 (13.52)
Fixed	-5.94 (0.484)	-5.05 (0.489)	-

Table 3.16: Variance Component Estimates for Full Labour Force Status with Employment as the Baseline Category: Repeated Panel Data

In the multinomial multilevel model (3.3) we are catering for the three categories of labour force status by fitting two related logistic multilevel models to the data. If we then also estimate the covariance between the parameters for the two non-baseline categories, we are tripling the number of parameters we are simultaneously trying to estimate compared to the simple logistic multilevel model (3.2) presented in the previous section. Given that Model (3.2) already has a complex variance decomposition structure (as highlighted in Figure 3.7) it is likely that the added

complexity represented in Model (3.3) will further reduce the convergence speed of the estimation procedure and potentially increase the uncertainty associated with interviewer effect estimates made under this model.

In this simple case where we are comparing the explanatory power of two competing models with differing variance decomposition structures, we can assess the appropriateness of Model (3.3) compared with Model (3.2) using the Bayesian Deviance Information Criteria (DIC) of Spiegelhalter *et al.* (2002). We can also compare the appropriateness of these variance decomposition models with the simple one level fixed effect model, $P^{(i)} = \Pr\left(y_q = i|\pi^{(i)}\right) = \frac{\exp(\pi^{(i)})}{1+\exp(\pi^{(i)})}$, $\pi^{(i)} = \mu$, in which there is no variance decomposition for each response, q.

Multilevel Model Type	$\bar{D}$	$D(\bar{\theta})$	pD	DIC
Multinomial (3.3)	4617.80	2595.88	2021.92	6639.73
Logistic (3.2)	3133.85	1755.50	1378.36	4512.21
Fixed effect only	24031.98	24030.97	1.00	24032.98

Table 3.17: Bayesian Deviance Information Criteria Comparison of Logistic and Multinomial Multilevel Models: Repeated Panel Data

In Table 3.17 the DICs for Model (3.3) and Model (3.2) are presented. The DIC is a statistic that contrasts the fit or adequacy of the model, as measured by the posterior mean deviance, $\bar{D}$, with the complexity of the model, as measured by the effective number of parameters, pD. It is important to penalize more complex models in this way as increasing the complexity of the model will never decrease the adequacy of the model fit, even though it will impact on the parsimony of the model. As the DIC statistic considers both the complexity and the adequacy of the model, it can be used to directly compare the appropriateness of the multinomial and logistic formulations. Lower DIC values indicate a more appropriate model for description of the data. The DIC and its application to model selection when estimating the interviewer effect will be discussed in more detail later in this chapter. Further information regarding the methodology underlying the DIC can be found in Spiegelhalter *et al.* (2002).

In this case we can see from Table 3.17 that decomposing the variance improves

the model fit as measured by the DIC. Both the multinomial and the logistic variance decomposition models have a lower DIC than the fixed effect only model and this indicates that it is important to decompose the variance in this case. We can also see that the multinomial variance decomposition model (3.3) has a higher DIC than the logistic variance decomposition model (3.2). This indicates that the greater complexity of the multinomial variance decomposition has a negative impact on the appropriateness of the model as measured by the DIC.

Thus the decision to use the multinomial variance decomposition model (3.3) for the estimation of the interviewer effect will ultimately depend on the aims of the analysis. If, for example, the difference between the interviewer effect for the three different categories of labour force status is important we may still choose to apply the multinomial variance decomposition model. On the other hand if the primary focus of the analysis is to estimate interviewer effects for monitoring purposes, such as monitoring the interviewer level residuals for particular interviewers, identifying questions and data items with high interviewer effects or determining the interviewer component of total survey error, the logistic variance decomposition model (3.2) may be appropriate. A further consideration for large datasets would also be the potential computing time required to estimate the interviewer effect under both of these models. Due to the increased complexity of Model (3.3) the logistic variance decomposition model (3.2) will generally be preferred for time critical analysis.

We have seen that the complexity of the multinomial variance decomposition model (3.3) may affect the stability of the estimates and inhibit convergence within a reasonable time frame. The following section will further address this concern by examining the performance of the multinomial variance decomposition model (3.3) for estimating variance components of a multinomial response variable simulated on the CURF structure with random effects of known magnitude.

MCMC Estimation of Multinomial Variance Decomposition Model on Simulated Dataset

We can provide further support for the estimates of the variance components of the multinomial variance decomposition model (3.3) by assessing MCMC estimates on a simulated multinomial response vector with known random effects. Here we apply the MCMC variance component estimates, excluding the covariance terms, in Table 3.16 to the CURF data structure to produce a simulated multinomial labour force status variable. An algorithm to simulate a multinomial response with known random effect sizes on the CURF structure can be found in Appendix A.

	Unemployed		NILF	
Level	Simulated	Estimated	Simulated	Estimated
Workload	0.1	0.10 (0.01)	0.9	3.12 (1.71)
Interviewer	0.15	0.05 (0.08)	0.12	0.48 (0.29)
CD	7.5	0.14 (0.15)	19.3	17.5 (5.39)
Household	7.1	0.44 (0.54)	60.9	63.9 (9.63)
Person	43.3	7.0 (1.06)	120.8	175.6(19.5)
Fixed	-5.7	-5.0 (0.23)	-4.7	-3.3 (0.51)

Table 3.18: Variance Component Estimates for Simulated Repeated Panel Labour Force Status with Employment as the Baseline Category

We can see in Table 3.18 that the MCMC variance component estimation procedure performs well at some levels but poorly at others. We would expect the MCMC estimates to be unbiased but due to the high correlation in our MCMC chain we have yet to converge to the appropriate posterior. In this simulation we know the *true* values of the variance components and can apply these *true* values as the initial values in our MCMC chain. However if we were trying to estimate the variance components without prior knowledge as to their magnitude the Brooks-Draper diagnostic indicates we would need to run the MCMC chain for approximately 27 million iterations to produce estimates for all levels to two significant figures. This chain length is determined by the household level variance component for the unemployed category which has the highest autocorrelation for this example and a corresponding Brooks-Draper value of 26,893,572.

Note that the estimates presented in Table 3.18 were calculated with a MCMC chain length of 450,000 iterations, which required approximately 20 hours of processing time on a Pentium III 1000 computer. Consequently this simulated example suggests that it may not be feasible to produce estimates of the variance components of labour force status using the multinomial variance decomposition model (3.3). There are a number of techniques that can be adopted to improve the efficiency of MCMC estimation of the parameters in large cross-classified multilevel models such as thinning, hierarchical centering (Gelfand *et al.* (1995)) and re-parameterization (Hill and Smith (1992)). An example of the practical application of these techniques can be found in Browne (2004). Given that Table 3.18 presents a simplified example dealing with fewer parameters than the real data presented in Table 3.16, it is likely that estimation of the interviewer effect under the multinomial variance decomposition model (3.3) will be too inefficient to apply in a time critical environment.

Summary

In this section we applied a multinomial variance decomposition model (3.3) to estimate the interviewer effect on labour force status. This model is an extension to the logistic variance decomposition model (3.2) considered in the previous section and introduces further complexity into the variance decomposition structure. We have seen that there is a cost associated with this increased complexity both in terms of speed and efficiency of estimates produced under model (3.3).

Although labour force status is an unordered multinomial data item, based on an objective DIC comparison, the logistic variance decomposition model (3.2) can be applied to more appropriately describe the data. Furthermore for the purposes of monitoring the interviewer effect, highlighting exceptional interviewers and producing an indication of the impact of the interviewer effect on total survey error, the logistic variance decomposition model (3.2) performs satisfactorily. Due to this and given its greater efficiency, the logistic variance decomposition model (3.2) will be applied exclusively in the remainder of this chapter for the estimation of the interviewer effect.

The following section will apply the logistic variance decomposition model (3.2) and consider the potential gain from including individual level covariates in the variance decomposition model. This may lead to more appropriate estimates of the interviewer effect as the covariates may help to explain some of the correlation between responses that might otherwise have been inappropriately attributed to the interviewer.

3.5 Controlling for Covariates

In the previous sections we have seen how to estimate the interviewer effect on the CURF dataset using the simple logistic variance decomposition model (3.2). This model was used to decompose the variation associated with a binary employment status variable into variance components relating to the classification levels in the CURF. In Chapter 1 we defined the interviewer effect as the impact on estimates due to the correlation between responses due to the presence and characteristics of the interviewer. Consequently application of the logistic variance decomposition model (3.2) would suggest that an estimate of the interviewer effect could be provided by the variance component at the interviewer level.

In practice, however, not all of the correlation between responses collected by an interviewer will be due to the presence and the characteristics of that interviewer. Some of this correlation may be due to the characteristics of the individual respondent, irrespective of the interviewer. For example retirees, possibly indicated by the respondent being above the age of 65, generally have a greater chance of being Not In the Labour Force (NILF). Similarly spatial areas that have characteristics in common may be more likely to contain respondents who are, in some way, more similar and who then report more similar responses. Therefore, in order to produce more appropriate estimates of the interviewer effect we need to control for some of the factors, not related to the interviewer, that may otherwise lead to correlation at the interviewer level. This can be done by extending the logistic variance decomposition model (3.2) to include covariates that may explain some of the correlation

between responses in the CURF.

3.5.1 Logistic Variance Decomposition Model Incorporating Covariates

We have already seen that there are a number of data items available in the CURF that may explain some of the correlation between responses for the employment status data item. For example at the individual level we have age, gender, country of birth and marital status. It would also be desirable to include higher level covariates, such as interviewer characteristics at the interviewer level and household characteristics at the household level, however this information is not available on the CURF. The remaining covariate that will be considered in this section is month, which has 4 categories; August, September, October and November 2001.

These covariates will be incorporated as fixed effect parameters added to the logistic variance decomposition model (3.2) discussed earlier. A selection of following fixed effect terms will be added to the model

- Age: A categorical covariate with 6 different categories, $Age1, ..., Age6$. Age category 1, $Age1$, corresponding to 15 to 24 year olds, was chosen as the baseline category. See Section 3.1.1 for more detail on the remaining age categories

- Gender: A categorical covariate with two different categories, $Gender1$ and $Gender2$. Category 1, $Gender1$, corresponding to the male gender, was chosen as the baseline category

- Country of Birth: A categorical covariate with three different categories, $Cob1$, $Cob2$ and $Cob3$. Category 1, $Cob1$, corresponding to respondents born in Australia, was chosen as the baseline category. See Section 3.1.1 for more detail on the other country of birth categories

- Marital Status: A categorical covariate with two different categories, $Marital1$ and $Marital2$. Category 1, $Marital1$, corresponding to living with a partner,

was chosen as the baseline category

- Time: A categorical covariate with four different categories, $Time1, ..., Time4$. Category 1, $Time1$, corresponding to August 2001, was chosen as the baseline category

Note that the effect of time will also be considered as a continuous covariate.

For the remainder of this chapter we will consider the potential gain from including a combination of these fixed effect terms as covariates in a logistic variance decomposition model. For comparison purposes all models considered in this section will apply the same variance decomposition structure as specified by the classification structure of the CURF highlighted in Figure 3.7. Given also the logistic link function the combination of fixed effect covariates can then be used to solely specify the model. For example if we add a fixed effect covariate for gender to the logistic variance decomposition model (3.2) specified previously, we would get the following model

$$\Pr(y_{ijlkmn} = 1|\pi_{jklmn}) = \frac{\exp(\pi_{jklmn})}{1 + \exp(\pi_{jklmn})} \tag{3.4}$$

where y is the employment status response and π can be decomposed into a number of variance components corresponding to the classification levels in the CURF

$$\pi_{jklmn} = \mu + \beta_1 Gender2_{jlkmn} + \omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn}$$

and

- μ is a fixed effect
- $Gender2$ is a fixed effect covariate indicating the effect on employment status of gender relative to the baseline (in this case male)
- The random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_{int})$ and $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$
- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively

- The classification structure of the data is highlighted in Figure 3.7

Similarly specifying the addition of fixed effect covariates for gender and country of birth would then lead to the following model

$$\Pr(y_{ijlkmn} = 1|\pi_{jklmn}) = \frac{\exp(\pi_{jklmn})}{1 + \exp(\pi_{jklmn})} \tag{3.5}$$

where π can again be decomposed into a number of variance components corresponding to the classification levels in the CURF

$$\pi_{jklmn} = \mu + \beta_1 Gender2_{jlkmn} + \beta_2 Cob2_{jlkmn} + \beta_3 Cob3_{jlkmn} + \omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn}$$

For the remainder of this section we will consider how the addition of fixed effect covariate terms into the variance decomposition model influences the interviewer effect estimate. In the following section model comparison techniques for determining the appropriate covariate model will be discussed. As has been demonstrated during this section the following logistic variance decomposition models will be described solely by the covariates included in the model.

3.5.2 Determining an Appropriate Variance Decomposition Model

The purpose of including covariates into the variance decomposition model is to control for some of the factors that may explain some of the variation at the interviewer level and would otherwise inflate or diminish the variance component estimates. This is recognized by Verbeke and Molenburghs (2000, p 122) who state

'*...the covariance structure itself may be of interest for understanding the random variation observed in the data. However, since it only explains the variability not explained by systematic trends, it is highly dependent on the specified mean structure.*'

However, including all possible covariates into the model, without considering the complexity and explanatory power of the model will also have implications for the appropriateness of the variance decomposition model.

We have already determined the complete variance decomposition structure for the CURF and have discussed that the consequence of ignoring any of the levels in this variance decomposition structure (following the work of Tranmer and Steel (2001); Hutchison and Healy (2001)) would be to inflate the interviewer effect estimate. Hence to produce appropriate interviewer effect estimates the variance decomposition structure highlighted in Figure 3.7 must be applied in the variance decomposition model.

We have also introduced the DIC of Spiegelhalter *et al.* (2002) that can be applied, under a Bayesian framework, to compare competing GLMMs. This statistic is somewhat controversial as we can see that alternative model comparison approaches such as determining the most appropriate model with approximate likelihood procedures and applying the likelihood ratio test to compare potential models (see for example Littell *et al.* (1996); Pickery and Loosveldt (2004)) will be inadequate in this case. This is because there is a strong bias in estimates produced with the approximate methods when estimating the interviewer effect associated with the CURF. Rodriguez and Goldman (2001) discussed how the bias in the approximate estimates is related to the magnitude of the variance components themselves. Consequently if the addition of the covariates explains some of the correlation and therefore reduces the magnitude of the variance components, then this will also potentially reduce the bias. In extreme cases the effect of the potential change in bias, rather than the actual model fit, could completely determine the most appropriate model under the approximate likelihood approach. The DIC is therefore a more appropriate model comparison tool in this case because it utilizes an appropriate estimation (and potentially unbiased) methodology to compare variance decomposition models.

We must note however that the DIC is a model comparison statistic that has been designed to evaluate the overall appropriateness of the model (see Spiegelhalter *et al.* (2002) for more information). In practical terms this will generally be a satisfactory condition. However when the primary focus of the analysis is to estimate a specific variance component such as the interviewer effect, it may be possible to specify more appropriate model comparison criteria although this is an area in which

more research is required. In this section the DIC will be applied to compare variance decomposition models with differing covariates. The variance decomposition structure presented in Figure 3.7 will be applied exclusively in this section.

Included Covariates	$\bar{D}$	$D(\bar{\theta})$	pD	DIC
Gender	3130	1760	1380	4510
Cob	3130	1760	1380	4510
Gender and Cob	3140	1760	1380	4520
Time	3140	1750	1380	4520
Age	3160	1770	1380	4540
Age, Gender, Cob and Marital	3180	1800	1400	4580
Marital	3180	1780	1410	4590
Age, Gender, Cob, Time and Marital	3220	1790	1430	4650
Time (as continuous variable)	3020	590	2430	5460
Covariate free model (3.2)	3130	1760	1380	4510

Table 3.19: Mean DIC for Variance Decomposition Model with Covariates: Repeated Panel Data

In Table 3.19 above mean DICs calculated over five separate MCMC runs initialized using different random seeds and initial values set to the final estimates are presented and rounded to three significant figures. Averaging over separate runs was performed as small differences in the DIC statistic may not be consistent over repeated MCMC runs. Note also that the relationship $\bar{D}+pD = DIC$ may not be preserved in Table 3.19 due to rounding. DICs were estimated for all possible combinations of fixed effect covariates. However only a selection of relevant models are presented in Table 3.19. We can see that on a DIC comparison there is no overall benefit from the inclusion of any of the covariates compared with the covariate free model (3.2).

The DIC is based on an estimate of the fit of the model, as measured by the posterior mean deviance, $\bar{D}$, and an estimate of the complexity of the model, considered by the effective number of parameters, pD. We can see in Table 3.19 that the DIC measures are similar for each fitted model. However, although the DIC indicates that there is no overall gain from incorporating covariates into the variance decomposition model, the complexity of the variance decomposition structure may inhibit our ability to determine the appropriateness of variance decomposition

models including covariates.

In traditional modelling the covariance structure seeks to explain the remaining variability in the data that cannot be explained by the fixed effects. Although specification of the fixed effects will influence the random effect estimates and vice versa, we can gain insight into the relevance of the covariates by considering the explanatory power of the covariates in a simple model with no variance decomposition structure.

As in Table 3.19, Table 3.20 refers to models via the included covariates. In all cases in Table 3.20 there is no variance decomposition structure incorporated along with the covariates to model the variation. For example if a model is referred to as containing the following covariates; age, gender, country of birth and marital status, the model can therefore be written as

$$\Pr(y_{ijlkmn} = 1|\pi_{jklmn}) = \frac{\exp(\pi_{jklmn})}{1 + \exp(\pi_{jklmn})} \tag{3.6}$$

where y is the employment status response and the probability of being not employed at π relates to the covariates as follows

$$\begin{aligned}\pi_{jklmn} = \mu &+ \beta_1 Gender2_{jklmn} + \beta_2 Marital2_{jklmn} + \beta_3 COB2_{jklmn} + \beta_4 COB3_{jklmn} \\ &+ \beta_5 Age2_{jklmn} + \beta_6 Age3_{jklmn} + \beta_7 Age4_{jklmn} + \beta_8 Age5_{jklmn} + \beta_9 Age6_{jklmn}\end{aligned}$$

and

- μ is a fixed effect
- $Gender2$ is a fixed effect covariate indicating the effect on employment status of gender relative to the baseline (in this case male)
- $Marital2$ is a fixed effect covariate indicating the effect on employment status of marital status relative to the baseline (in this case living with partner)
- $COB2$ and $COB3$ are fixed effect covariates indicating the effect on employment status of country of birth relative to the baseline (in this case born in Australia)

- *Age*2, *Age*3, *Age*4, *Age*5 and *Age*6 are fixed effect covariates indicating the effect on employment status of age of respondent relative to the baseline (in this case 15 to 24 year olds)
- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively

Then by considering the potential gain from incorporating the covariates in the absence of the complex variance decomposition structure, we can determine more easily which covariates may be included to explain some of the variation in employment status on the CURF.

Included Covariates	$\bar{D}$	$D(\bar{\theta})$	pD	DIC
Age, Gender, Cob and Marital	18300	18300	10	18300
Age, Gender, Cob, Time and Marital	18300	18300	13	18300
Age, Gender, Time and Marital	18400	18400	11	18400
Age, Gender, Time and Cob	18500	18400	12	18500
Age	19100	19100	6	19100
Gender	23600	23600	2	23600
Marital	23900	23800	2	23800
Cob	23900	23900	3	23900
Time	24000	24000	4	24000
Constant only Model	24000	24000	1	24000

Table 3.20: DIC Statistics for Selected Covariate and Residual only Models: Repeated Panel Data

We can see in Table 3.20 that the estimates of the effective number of parameters align with the actual number of parameters in the models. For example in Model (3.6) we can see that there are 10 parameters to be estimated, the mean, μ, 5 for age, 2 for country of birth and 1 each for marital status and gender. In Table 3.20 the estimated effective number of parameters for Model (3.6), referred to as Age, Gender, Cob and Marital, is also 10. This indicates that for the simple covariate models, without variance decomposition, we are getting appropriate estimates of the effective number of parameters, pD, which provides support for the DIC statistic in

this case. Consequently due to the decreased complexity of these models compared with the variance decomposition models presented in Table 3.19 we might put more weight on the DIC comparisons presented in Table 3.20 to determine the covariates that should be incorporated in the final variance decomposition model. On a DIC comparison basis, Table 3.20 indicates that Model (3.6) would be preferred, suggesting that covariates for age, gender, country of birth and marital status all can be incorporated to explain some of the variation in employment status on the CURF. Of the tested covariates, only time did not lead to an improvement in the model fit. This makes intuitive sense as the reference period of the CURF (August to November 2001) was one of stable, but slow employment growth. This can be seen in the CURF as the unemployment rate for August 2001 of 5.5% is the same as as the unemployment rate in November 2001. We have also seen in Table 3.9 that the majority of individuals in the CURF did not change their employment status in any two consecutive months and this suggests that there was not a large change over time in employment status. This is further borne out by the parameter estimates produced by models incorporating time as a covariate, as the time parameter is not significant in all months.

The potential inclusion of the fixed effect parameters for age, gender, marital status and country of birth is intuitively appealing. This is because it would be expected that these demographic factors are related to the employment status of an individual. For example we would generally expect persons above the age of 65, i.e. in age category 6, would have a lower probability of being employed as they are more likely to be retired. The analysis of the potential gain from including these covariates ignoring the variance decomposition structure (as presented in Table 3.20) is therefore intuitively appealing. However due to the interaction between the estimation of the random and fixed effect parameters we cannot blindly apply this analysis to determine the fixed effect covariates that should be included in the variance decomposition model. We also cannot solely rely on the DIC model comparison of Table 3.19 to determine the fixed effect covariates that should be included in this model due to the complexity of the fitted models and the variability

inherent in the estimated components of the DIC.

As the final form of the variance decomposition structure has been determined earlier and we have seen the consequences of altering this structure would be to potentially inflate the interviewer effect estimate, we only have to determine the fixed effect covariates that should be included in the variance decomposition model. Table 3.20 suggests therefore that we might also want to consider the following variance decomposition model

$$\Pr(y_{ijlkmn} = 1|\pi_{jklmn}) = \frac{\exp(\pi_{jklmn})}{1 + \exp(\pi_{jklmn})} \tag{3.7}$$

where y is the employment status response and the probability of being not employed, π relates to the covariates as follows.

$$\begin{gathered}\pi_{jklmn} = \mu + \beta_1 Gender2_{jklmn} + \beta_2 Marital2_{jklmn} + \beta_3 COB2_{jklmn} + \beta_4 COB3_{jklmn} \\ +\beta_5 Age2_{jklmn} + \beta_6 Age3_{jklmn} + \beta_7 Age4_{jklmn} + \beta_8 Age5_{jklmn} + \beta_9 Age6_{jklmn} \\ +\omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn}\end{gathered}$$

and

- μ is a fixed effect
- $Gender2$ is a fixed effect covariate indicating the effect on employment status of gender relative to the baseline (in this case male)
- $Marital2$ is a fixed effect covariate indicating the effect on employment status of marital status relative to the baseline (in this case living with partner)
- $COB2$ and $COB3$ are fixed effect covariates indicating the effect on employment status of country of birth relative to the baseline (in this case born in Australia)
- $Age2$, $Age3$, $Age4$, $Age5$ and $Age6$ are fixed effect covariates indicating the effect on employment status of age of respondent relative to the baseline (in this case 15 to 24 year olds)
- The random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_{int})$ and $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$

- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively
- The classification structure of the data is highlighted in Figure 3.7

Hence based on our analysis of the potential gain from incorporating covariates we have three suggested models that we can compare,

1. The simple variance decomposition model (3.2) which contains no covariates. This model is the simplest model and also performs well in a DIC comparison in Table 3.19.

2. A variance decomposition model (3.4) incorporating gender as a covariate. This is the model marginally preferred under the DIC model comparison presented in Table 3.19.

3. A variance decomposition model (3.7) incorporating the covariates determined in Table 3.20, i.e. covariates for age, gender, marital status and country of birth.

We have already seen DIC comparison statistics for these three models. In the table below we now compare the parameter estimates for the variance decomposition models with differing covariates.

Variance Component	Model (3.2)	Model (3.4)	Model (3.7)
Workload	3.12 (2.27)	3.74 (3.61)	2.40 (1.23)
Interviewer	0.10 (0.09)	0.10 (0.09)	0.07 (0.08)
CD	16.02 (4.04)	15.43 (3.84)	4.88 (1.67)
Household	84.18 (10.11)	87.75 (10.16)	36.28 (4.93)
Person	107.94 (10.99)	97.44 (10.04)	67.76 (7.23)

Table 3.21: Variance Component Estimates for Repeated Panel Variance Decomposition Models

We can see in Table 3.21 that incorporating covariates into Model (3.7) has reduced the magnitude of all of the variance component estimates in comparison with

Model (3.2). Examination of the diagnostics for each of these variance components indicates that the posterior is normally distributed for all variance components except for the interviewer effect. Hence we can interpret the Bayesian equivalences as a standard error and see that there has been a significant reduction, at the 95% level in the CD, household and person level variance components. In comparison there is no significant reduction in the workload level variance component and the interviewer effect is also not significantly reduced. Consequently we can see that, for the purposes of estimating the interviewer effect, there has been a noticeable change in the relevance of the interviewer effect (as compared to the total unexplained variation) when covariates are included in the variance decomposition model. This can be seen in Table 3.21 as the estimated ρ_{int} is 31% larger for Model (3.7) than in Model (3.2).

In Table 3.21 it appears that the inclusion of a number of person level covariates has had a significant impact on explaining the correlation at the person, household and CD levels. This is intuitively appealing as these are the levels most affected by the characteristics of individual people in the CURF and over larger groupings, such as the workload and interviewer levels, we might expect the average characteristics of respondents to be similar. The inclusion of interviewer level covariates may have a much larger impact on the interviewer effect estimate. However we do not have this information available on the CURF. Although the inclusion of person level covariates does not seem to have had a large impact on the interviewer effect estimate we can again examine the interviewer level residuals to explore whether the reduced unexplained variation in Model (3.7) allows us to better distinguish interviewers.

We can see in Figure 3.11 that we still do not have enough information to be able to distinguish between interviewer residuals. Despite some minor changes to the interviewer residual estimates this graph is strongly reminiscent of Figure 3.10. Hence we can see that the inclusion of the person level covariates has not had much impact on the interviewer effect estimate.

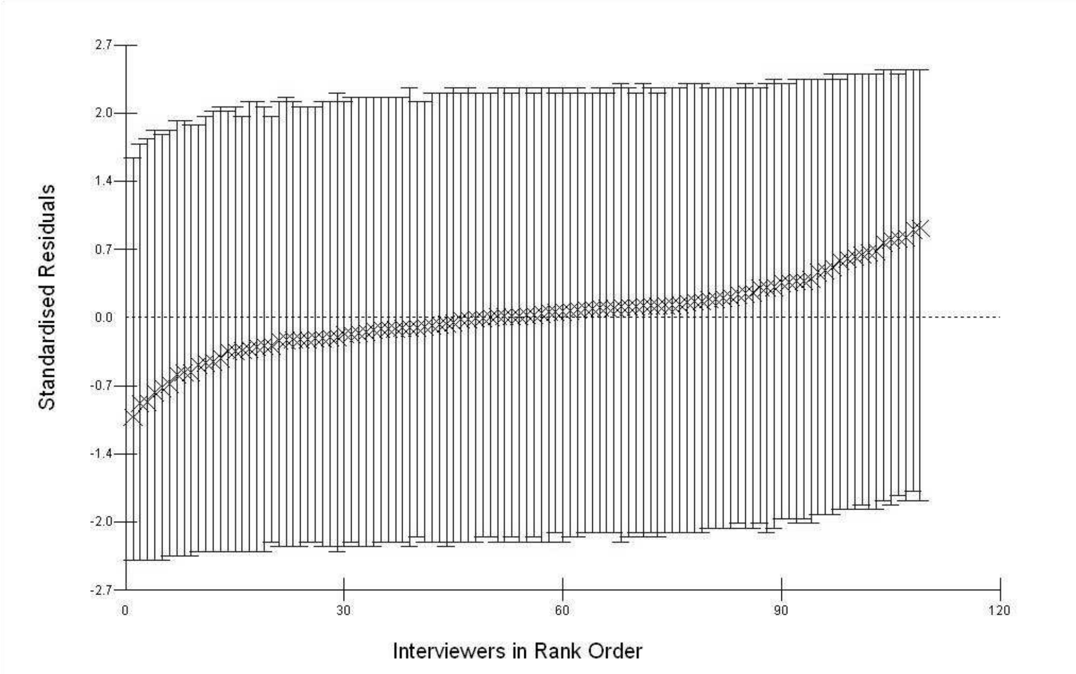

Figure 3.11: Interviewer Level Residuals for Repeated Panel Logistic Variance Decomposition Model (3.7)

3.5.3 Interpretation of Results

Although we have seen that inclusion of a number of person level covariates in the variance decomposition model does not have much impact on the interviewer effect estimates, the inclusion of the covariates does effect the interpretation of the overall impact of the interviewer on results derived from the CURF. We have seen that on a DIC comparison basis, there is little to choose between any of the variance decomposition models incorporating covariates. On the other hand we have seen that the inclusion of age, gender, country of birth and marital status as covariates leads to a significant reduction in the unexplained variation in Model (3.7). We also saw in Table 3.20 that all of these covariates appear to be relevant in explaining the variation in employment status.

Generally person level covariates such as age and gender will have some relationship to the employment status of individuals, for example, levels of employment will greatly decrease above retirement age. Consequently we would conclude that the covariate inclusive variance decomposition model (3.7) is the most appropriate for

estimating the interviewer effect.

In Table 3.21 we estimated the interviewer effect as a variance component on employment status and saw that it was small in relation to the remaining variance components. This agreed with our initial observations in Table 3.9 which also suggested a small interviewer effect. However, given that the remaining variance components are spatial effects and of intrinsic interest to the data analyst, this does not mean that the interviewer effect does not have an important impact on employment status estimates derived from the CURF. We can consider the overall impact of the interviewer effect on employment status by calculating the intraclass correlation coefficient introduced in Chapter 2.

Snijders and Bosker (1999, p 224) discuss how in the logistic variance decomposition model the level one residuals follow a logistic distribution and therefore have an implied variance of $\frac{\pi^2}{3} = 3.29$. Based on this we can say that the intra-interviewer correlation coefficient, ρ_{int}, under variance decomposition model (3.7) will be 6.1×10^{-4} or approximately 0.0006. Which according to result (2.3) would lead to a 14.1% increase in the measurement variance component of the total survey error. Note that the variation in the number of respondents allocated to interviewers also has an impact. If each of our 109 interviewers collected data from $\frac{18264}{109}$ or approximately 167 respondents, then this would have lead to only a 10.1% increase in the measurement variance component of the total survey error.

The intra-interviewer correlation coefficient, ρ_{int}, in this case is small compared with the majority of studies of the interviewer effect, e.g.

- Hansen and Marks (1958) who found that $\rho_{int} = 0.011$
- Kish (1962) who found that $\rho_{int} = 0.02$
- Felligi (1964) who found that $\rho_{int} = 0.008$
- Freeman and Butler (1976) who found that $\rho_{int} = 0.036$
- O'Muirceartaigh and Wiggins (1981) who found that $\rho_{int} = 0.02$
- Groves and Magilavy (1986) who found that $\rho_{int} = 0.009$

- O'Muircheartaigh and Campanelli (1998) who found that $\rho_{int} = 0.075$

However, these studies generally focused on subjective question items in which the interviewer effect was expected to be large. In comparison we would expect for relatively objective data items such as employment status on the CURF that the intra-interviewer correlation coefficient would be smaller than in the above studies. This was indeed the case, however we have seen that when interviewers deal with a large number of respondents, even a small intra-interviewer correlation coefficient can lead to an increase in the measurement variance component of total survey error.

We can now expand there results to produce estimates of the total variance. Let p be the proportion of adults in the population who are employed, then for a simple random sample we can estimate this proportion from the CURF, i.e. $\hat{p} = \frac{n_{emp}}{n}$ where n_{emp} is the number of respondents employed in the CURF. Then $\hat{p} = 0.632$ and the usual estimate, $V\hat{a}r(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1} = 1.273 \times 10^{-5}$ ignores the impact of the interviewer effect. We can then apply (2.10) to produce an estimate of the total variance of the estimated proportion employed in the population that appropriately considers the interviewer effect. In this case the estimated total variance is 1.454×10^{-5}, 14% greater than that predicted by the usual estimate. Although respondents in the CURF have not been selected via simple random sampling, this simple demonstration clearly indicates the impact of interviewer effects on estimates derived from household surveys.

Summary

We have seen that it is important to incorporate fixed effect covariates to control for unexplained correlation that could otherwise inflate or shrink the interviewer effect estimate. In practice there is currently no clear consensus regarding how to determine which covariates should be included in the model, as model search techniques utilizing approximate estimation methodologies are biased while Bayesian techniques for model comparison remain controversial and are very computationally intensive. Furthermore, the focus of current techniques is generally on determining the most appropriate fixed effect structure rather than choosing the optimal model

for the estimation of a specific variance component. It is only after we choose an appropriate variance decomposition structure for the estimation of the interviewer effect that the choice of fixed effect covariates can be determined. More research is needed into how to simultaneously determine the most appropriate variance decomposition structure and covariates to be included in variance decomposition models for the estimation of the interviewer effect.

3.6 Discussion

In this chapter we have demonstrated how techniques for estimating the interviewer effect can be extended in practice to include the longitudinal information available in repeated panel surveys. With reference to a CURF drawn from the LFS we have shown that incorporating the longitudinal information as an extra level in the classification structure of the data can lead to a large increase in the effective degree of interpenetration that we can associate with the survey. We would expect that increased degrees of partial interpenetration will generally lead to improved estimates of the interviewer effect and this chapter demonstrated how the interviewer effect could be estimated in a partially interpenetrated survey.

In practice surveys are generally neither fully interpenetrated nor fully confounded unless there is large expenditure dedicated to producing a fully interpenetrated design. Consequently, the ability to consider partially interpenetrated surveys is an important advance in the practical estimation of the interviewer effect. The following chapters will explore the implications of partial interpenetration in more detail, defining partial interpenetration for the first time and determining the effect, in general, of partial interpenetration on interviewer effect estimates.

In this chapter we have also demonstrated some of the practical issues involved with estimating the interviewer effect. Rodriguez and Goldman (2001) suggested that for non-linear models the bias from estimating variance components using approximate techniques such as MQL or PQL is related to the number of level one units within the level two units. In household surveys it will generally be the case

that there are only a small number of responding persons within each household and hence we expect the bias in the variance component estimates to be large on non-linear data items collected via household surveys. We showed that there is a large bias in the estimates of variance components of the employment status variable on the CURF and that exact techniques such as MCMC can be used to produce more appropriate estimates. This is the first study to consider the potential bias inherent in estimation techniques when dealing with the interviewer effect and to actively apply these techniques to a large scale survey.

This chapter has demonstrated how current techniques can be extended to estimate the interviewer effect in practice and indicated some of the advantages and limitations of the current techniques. In particular, estimation methodologies are becoming much more flexible for considering the complex classification structures of large scale surveys and consequently more sophisticated variance decomposition models can now be considered. On the other hand the theory behind estimating the interviewer effect must be expanded. A new comprehensive definition of interpenetration, incorporating partial interpenetration must be compiled and model comparison tools for determining appropriate variance decomposition models for the estimation of the interviewer effect have still to be developed.

In summary regular estimation of the interviewer effect is an important quality monitoring tool, indicating questions in a survey that may be causing confusion and interviewers who are performing unusually. This will aid in the monitoring of surveys, lead to more efficiently targeted training regimes and allow designers to more effectively allocate resources to improve the quality of surveys.

Chapter 4

Defining Partial Interpenetration

We have seen in Chapter 3 how modern computing techniques can be applied to estimate the interviewer effect provided interviewer and spatial effects are not fully confounded. When fully interpenetrated designs have not been implemented these techniques can be applied to produce an estimate of the interviewer effect.

This chapter will revisit the traditional definition of interpenetration and comprehensively define the concept of partial interpenetration for the first time. We will then demonstrate how to determine the degree of interpenetration for a survey, either during the analysis or the design phase.

In practice most large-scale surveys are not designed to be interpenetrating because of the high costs involved. These surveys are generally run on a strict budget which precludes the use of fully interpenetrated designs. Consequently the introduction of partially interpenetration will allow statisticians to evaluate both the costs and the benefits of small increases in the degree of interpenetration associated with a survey, ultimately leading to more cost-effective estimates of the interviewer effect.

4.1 Defining Interpenetration

This section will explore how design issues relate to the confounding of variance components such as the interviewer effect. We can consider this from two perspectives,

- The survey methodologist who wants to design a survey that is not fully interpenetrated but gives as good estimates of the interviewer effect as possible (based on a constraint - e.g. cost) and;
- The data analyst who is given output data from a survey and who wants to assess the level of interpenetration and the implications this has on estimation of the interviewer effect.

Hence there are two questions,

- What is the degree of interpenetration implied by a given survey design?
- What effect will this level of interpenetration have on our estimate of the interviewer effect?

Current research for estimating the interviewer effect has relied on fully interpenetrated survey designs, while implicitly considering interviewer and spatial effects to be completely confounded if the design is not fully interpenetrated. We now extend this scenario to consider the appropriateness of interviewer effect estimates when dealing with surveys that are neither fully interpenetrated nor fully confounded under the traditional definition of interpenetration.

4.1.1 The Traditional Definition of Interpenetration

In order to incorporate the concept of partial, or incomplete, interpenetration we need to rigorously define both confounding and interpenetration. Our initial discussion of interpenetration relied upon a definition similar to that of Mahalanobis (1946).

Definition 4.1.1. *Interviewer interpenetration within spatial regions occurs when at least 2 interviewers are randomly allocated to each of the spatial regions.*

As early techniques for estimating the interviewer effect required active manipulation of the survey design, any non-interpenetrated designs were therefore considered to be confounded. We saw in Chapter 3 that we can make estimates of

the interviewer effect when we do not have full interpenetration, provided at least some of the spatial zones are interpenetrated, although we have seen in Table 3.11 that the variability of the interviewer effect estimates increased as the degree of interpenetration decreased.

Traditional definitions of interpenetration such as 4.1.1 have therefore become limiting as we can now use modern computing techniques to estimate the interviewer effect even when we do not have full interpenetration. Thus when considering the impact of partial or incomplete interpenetration on estimation of the interviewer effect a more comprehensive definition of interpenetration must be adopted. The following section will briefly review our GLMM framework for consideration of the interviewer effect and discuss concepts that definitions of interpenetration attempt to capture. A more comprehensive definition of interpenetration which incorporates partial interpenetration will then be introduced and methods for determining the degree of interpenetration based on this definition discussed.

4.1.2 A GLMM Framework for Defining Interpenetration

This section presents a brief review of GLMMs and how they can be utilized to consider the interviewer effect. More detail on GLMMs and further examples can be found in Chapter 2.

The general form of the GLMM was introduced as Model (2.12), where $\eta = \mathbf{X}\beta + \mathbf{Zu}$ and $\mathbf{Z}$ is a $(n \times t)$ known matrix, of which all elements are either 0 or 1, that serves as the design matrix for the random effects as specified in the $(t \times 1)$ random effect vector, $\mathbf{u}$. It is clear that the design matrix for the random effects, $\mathbf{Z}$, can be used to indicate how interviewers are allocated to spatial regions.

For example if we had six observations, which had been divided into two spatial

zones and allocated to two interviewers, we could represent this system as follows

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \end{bmatrix} \tag{4.1}$$

We can see that the design matrix, $\mathbf{Z}$, can be used to indicate how interviewers have been allocated to spatial regions. This is because the pth row of $\mathbf{Z}$ indicates to which interviewer and spatial zone the pth observation in our dataset belongs. For example in (4.1) the first row of $\mathbf{Z}$ corresponds to the first of the six observations in our dataset and we can see that this observation was collected by interviewer 1 in spatial zone 1.

Then we can also see how the design matrix, $\mathbf{Z}$, relates to the variance covariance matrix, $\mathbf{V}$, from which the variance components such as the interviewer effect are estimated through the expansion of $\mathbf{V}$ given in McCulloch and Searle (2001, pg 161)

$$\mathbf{V} = \sigma_\varepsilon^2 \mathbf{I_n} + \sum\nolimits_{i=1}^{t} \mathbf{Z_i} \mathbf{Z_i^T} \sigma_i^2 \tag{4.2}$$

where the design matrix, $\mathbf{Z}$, has been partitioned into t $(n \times 1)$ column vectors, $\mathbf{Z_i}$, $i = 1, ..., t$ such that

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_i & \cdots & \mathbf{Z}_t \end{bmatrix}$$

and σ_i^2 is the random effect associated with the ith row of $\mathbf{u}$. Note that we have already assumed in Chapter 2 that the spatial and interviewer effects are independent and random effects for all interviewers are distributed normally with zero mean and a variance of σ_{int}^2, i.e. in (4.1) for all interviewers, $j = 1, 2$, $Interviewer_j \sim iid(0, \sigma_{int}^2)$. Similarly for all of the spatial zones in (4.1), $k = 1, 2$, $Zone_k \sim iid(0, \sigma_{wk}^2)$. Thus the design matrix, $\mathbf{Z}$, can be partitioned into t $(n \times 1)$ column vectors, $\mathbf{Z_i}$, $i = 1, ..., t$ and the structure of the design matrix $\mathbf{Z}$ will therefore directly determine the elements of $\mathbf{V}$.

We can then also expand expression (4.2) to consider combinations of the columns of the design matrix, $\mathbf{Z}$, by recognising that

$$\begin{aligned} \mathbf{Z}\mathbf{Z}^{\mathbf{T}} &= \mathbf{Z}_1\mathbf{Z}_1^{\mathbf{T}} + \ldots + \mathbf{Z_i}\mathbf{Z_i^T} + \ldots + \mathbf{Z_t}\mathbf{Z_t^T} \\ &= \sum_{i=1}^{t} \mathbf{Z_i}\mathbf{Z_i^T} \end{aligned} \tag{4.3}$$

and we can then use (4.3) to simultaneously consider all columns of the design matrix, $\mathbf{Z}$, that relate to specific variance components such as the interviewer effect.

Given that the design matrix, $\mathbf{Z}$, is a $(n \times t)$ matrix that pre-multiplies the $(t \times 1)$ random effect vector, $\mathbf{u}$, we can see that each column of $\mathbf{Z}$ will relate to only one random effect in $\mathbf{u}$. In (4.1) therefore the first column of the design matrix $\mathbf{Z_1}$ relates to the first random effect specified in $\mathbf{u}$, i.e. the first spatial zone, $Zone_1$. Applying (4.3) we can then extend (4.2) to consider all of the random effect terms in $\mathbf{u}$ that relate to the interviewer effect by simultaneously considering all columns of $\mathbf{Z}$ that relate to interviewers. For example in (4.1) all of the design information regarding the interviewer effect is captured in the third and fourth columns of $\mathbf{Z}$ as these are the columns that relate to the interviewer terms in the random effect vector, $\mathbf{u}$.

Thus if we define $\mathbf{Z_{wk}}$ as a matrix comprised of the columns of $\mathbf{Z}$ that relate to the spatial zones or workloads and $\mathbf{Z_{int}}$ as a matrix comprised of the columns of $\mathbf{Z}$ that relate to the interviewers, we can see in (4.1) that $\mathbf{Z_{wk}}$ is formed by the first 2 columns of $\mathbf{Z}$ and $\mathbf{Z_{int}}$ is given by the final 2 columns of $\mathbf{Z}$. Note that in (4.1) it turns out that $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ happen to be equal,

$$\mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{Z_{int}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

We can combine this with our expression for the variance covariance matrix in the

simple 2 variance component model presented in (4.1) to say that

$$\mathbf{V} = \sigma_{\varepsilon}^{2}\mathbf{I_n} + \sigma_{int}^{2}\mathbf{Z_{int}}\mathbf{Z_{int}^{T}} + \sigma_{\mathbf{wk}}^{\mathbf{2}}\mathbf{Z_{wk}}\mathbf{Z_{wk}^{T}} \tag{4.4}$$

then it is clear under the GLMM that the impact of variance components such as the interviewer effect on the variance covariance matrix, $\mathbf{V}$ will be solely determined by the columns of the random effects design matrix that relate to that variance component, such as $\mathbf{Z_{int}}$ in the case of the interviewer effect.

4.1.3 Conceptualizing Interpenetration

This section will discuss the aims of interpenetration and how it allows us to estimate the interviewer effect. These concepts will be adapted to our GLMM framework and the implications for defining interpenetration discussed.

In essence interpenetration attempts to guarantee that the interviewer effect term will appear separately from the spatial effect terms in the variance covariance matrix, allowing separate estimation of these variance components. Our initial Definition 4.1.1 touched on two central concepts that relate to the degree of interpenetration of two classification structures, which are in this case the interviewer classification structure implied by $\mathbf{Z_{int}}$ and the spatial workload classification structure implied by $\mathbf{Z_{wk}}$

- Intergroup interpenetration: How many groups (i.e. spatial zones or interviewers) in either $\mathbf{Z_{int}}$ (or $\mathbf{Z_{wk}}$) contain some degree of interpenetration? I.e. how may spatial zones contain more than one interviewer?

- Intragroup interpenetration: How many observations within a single spatial zone or workload of $\mathbf{Z_{wk}}$ belong to different interviewers in $\mathbf{Z_{int}}$?

We can see that in order to estimate the interviewer effect, we want to maximize the number of elements of $\mathbf{V}$ in which the interviewer and spatial effects appear separately and this will generally require high levels of both intergroup and intragroup interpenetration. We can therefore produce a more rigorous definition of full

interpenetration by simultaneously maximizing the degree of both intergroup and intragroup interpenetration.

By way of example we can see that in (4.1) we had low intergroup interpenetration, because none of the spatial zones are interpenetrated and low intragroup interpenetration because there is only one interviewer in each spatial zone. If we then rearrange the design so that we have full interpenetration according to the traditional Definition 4.1.1 we might have

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \end{bmatrix} \tag{4.5}$$

In (4.5) we have full intergroup interpenetration because all of the spatial zones or workloads are interpenetrated and contain more than one interviewer. On the other hand we can see that we do not have full intragroup interpenetration because some of the observations within each spatial zone are collected by the same interviewer. We can see the implications of not having full intragroup interpenetration if we apply Equation (4.4) to examine the variance covariance matrix associated with this design

$$\mathbf{V} = \begin{bmatrix} \sigma_C^2+\sigma_\varepsilon^2 & \sigma_C^2 & \sigma_{wk}^2 & 0 & \sigma_{\text{int}}^2 & 0 \\ \sigma_C^2 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_{wk}^2 & 0 & \sigma_{\text{int}}^2 & 0 \\ \sigma_{wk}^2 & \sigma_{wk}^2 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_{\text{int}}^2 & 0 & \sigma_{\text{int}}^2 \\ 0 & 0 & \sigma_{\text{int}}^2 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_{wk}^2 & \sigma_C^2 \\ \sigma_{\text{int}}^2 & \sigma_{\text{int}}^2 & 0 & \sigma_{wk}^2 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_{wk}^2 \\ 0 & 0 & \sigma_{\text{int}}^2 & \sigma_C^2 & \sigma_{wk}^2 & \sigma_C^2+\sigma_\varepsilon^2 \end{bmatrix} \tag{4.6}$$

where $\sigma_C^2 = \sigma_{int}^2 + \sigma_{wk}^2$. In (4.6) there is still some confounding remaining in $\mathbf{V}$ because the combined effect σ_C^2 is still evident in some elements of $\mathbf{V}$. If we want to minimize the number of elements in $\mathbf{V}$ in which the interviewer and spatial effects

are confounded (i.e. do not appear separately in $\mathbf{V}$) and hence maximize the number of elements of $\mathbf{V}$ from which we can estimate the interviewer effect, we must also maximize the intragroup interpenetration. A design with both full intergroup and full intragroup interpenetration can be seen in (4.7) following

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \\ Interviewer_3 \end{bmatrix} \tag{4.7}$$

We can see in (4.7) that we have full intragroup interpenetration in this case because all of the observations within a single spatial zone are collected by different interviewers. This will then lead to a variance covariance matrix in which there is no confounding between the interviewer and spatial effects, in the off-diagonal elements of $\mathbf{V}$. This can be seen more clearly in (4.8) following

$$\mathbf{V} = \begin{bmatrix} \sigma_C^2 + \sigma_\varepsilon^2 & \sigma_{wk}^2 & \sigma_{wk}^2 & \sigma_{\text{int}}^2 & 0 & 0 \\ \sigma_{wk}^2 & \sigma_C^2 + \sigma_\varepsilon^2 & \sigma_{wk}^2 & 0 & \sigma_{\text{int}}^2 & 0 \\ \sigma_{wk}^2 & \sigma_{wk}^2 & \sigma_C^2 + \sigma_\varepsilon^2 & 0 & 0 & \sigma_{\text{int}}^2 \\ \sigma_{\text{int}}^2 & 0 & 0 & \sigma_C^2 + \sigma_\varepsilon^2 & \sigma_{wk}^2 & \sigma_{wk}^2 \\ 0 & \sigma_{\text{int}}^2 & 0 & \sigma_{wk}^2 & \sigma_C^2 + \sigma_\varepsilon^2 & \sigma_{wk}^2 \\ 0 & 0 & \sigma_{\text{int}}^2 & \sigma_{wk}^2 & \sigma_{wk}^2 & \sigma_C^2 + \sigma_\varepsilon^2 \end{bmatrix} \tag{4.8}$$

By incorporating full intragroup as well as full intergroup interpenetration in (4.7) we have removed any confounding in the off diagonal elements of $\mathbf{V}$. In general this will then lead to higher numbers of elements in $\mathbf{V}$ from which we can estimate the interviewer effect.

In contrast when we have full confounding, we never observe any responses collected by interviewers independently from spatial zones meaning that we cannot separately estimate these variance components without making some further restricting assumptions, such as regarding the form of the spatial effect. This can be

seen in our confounded example (4.1) in which the variance covariance matrix was

$$\mathbf{V} = \begin{bmatrix} \sigma_C^2+\sigma_\varepsilon^2 & \sigma_C^2 & \sigma_C^2 & 0 & 0 & 0 \\ \sigma_C^2 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_C^2 & 0 & 0 & 0 \\ \sigma_C^2 & \sigma_C^2 & \sigma_C^2+\sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_C^2 & \sigma_C^2 \\ 0 & 0 & 0 & \sigma_C^2 & \sigma_C^2+\sigma_\varepsilon^2 & \sigma_C^2 \\ 0 & 0 & 0 & \sigma_C^2 & \sigma_C^2 & \sigma_C^2+\sigma_\varepsilon^2 \end{bmatrix} \tag{4.9}$$

In (4.9) we cannot estimate the interviewer effect from $\mathbf{V}$ because the interviewer effect term never appears separately in $\mathbf{V}$ but only the combined effect, $\sigma_C^2 = \sigma_{int}^2 + \sigma_{wk}^2$. Thus it is clear that interviewer effects are fully confounded with spatial effects when variance components for the spatial effect, σ_{wk}^2, and the interviewer effect, σ_{int}^2, do not appear separately in any element v_{ij} of $\mathbf{V}$.

Now that we have considered the concepts that definitions of interpenetration and confounding attempt to capture, we can produce a more comprehensive definition of both interpenetration and confounding that will allow us to consider partially interpenetrated designs. The next section will examine confounding in more detail.

4.1.4 Defining Confounding in Detail

The concept we are attempting to capture when we consider full confounding is that we never observe interviewers independently from spatial zones, leading to full confounding in the variance covariance matrix $\mathbf{V}$. Thus when there is full confounding we never separate the interviewer effect σ_{int}^2 from the spatial effect σ_{wk}^2 and can only estimate the combined effect $\sigma_C^2 = \sigma_{int}^2 + \sigma_{wk}^2$. Consequently we can estimate σ_C^2 under the fully confounded scenario, but are unable to estimate either σ_{int}^2 or σ_{wk}^2.

We can therefore define full confounding as occurring when

Definition 4.1.2. *Interviewer effects are fully confounded with spatial effects when each interviewer completely enumerates an entire spatial region and each interviewer is allocated to only that spatial region.*

We can then also combine Definition 4.1.2 with Equation (4.4) to determine confounding based solely on the design matrix, $\mathbf{Z}$.

Definition 4.1.3. *Let $\mathbf{Z_{wk}}$ refer to the columns of the design matrix, $\mathbf{Z}$, that relate to spatial zones and let $\mathbf{Z_{int}}$ refer to the columns of $\mathbf{Z}$ that relate to the interviewer. Then interviewers effects are fully confounded with spatial effects if* $\mathbf{Z_{int}}\mathbf{Z_{int}^T}= \mathbf{Z_{wk}}\mathbf{Z_{wk}^T}$.

Definition 4.1.3 follows directly from (4.4) as if $\mathbf{Z_{int}}\mathbf{Z_{int}^T}= \mathbf{Z_{wk}}\mathbf{Z_{wk}^T}$ then $\mathbf{V} = \sigma^2_\varepsilon \mathbf{I}_n + (\sigma^2_{\text{int}} + \sigma^2_{wk})\mathbf{Z}_{\text{int}}\mathbf{Z}^{\mathbf{T}}_{\text{int}} = \sigma^2_\varepsilon \mathbf{I}_n + (\sigma^2_{\text{int}} + \sigma^2_{wk})\mathbf{Z}_{\text{wk}}\mathbf{Z}^{\mathbf{T}}_{\text{wk}}$ and hence we cannot identify any difference between the variance components relating to the interviewers or the spatial areas. This means the interviewer effect term does not appear separately from the spatial effect term in the variance covariance matrix, i.e. rather than being able to estimate either σ^2_{int} or σ^2_{int} only the combined variance component $\sigma^2_C = \sigma^2_{int} + \sigma^2_{wk}$ will appear in $\mathbf{V}$. Consequently we will be able to estimate the combined variance component σ^2_C but not separate out the interviewer effect.

By way of example consider (4.1) in which $\mathbf{Z_{int}}\mathbf{Z_{int}^T}= \mathbf{Z_{wk}}\mathbf{Z_{wk}^T}$ and recall the associated variance covariance matrix in (4.9). In (4.9) we can see that the interviewer effect term never appears separately from the spatial effect, and we can only estimate the combined effect σ^2_C. Hence we cannot estimate the interviewer effect in this fully confounded scenario.

Under the classical definition of interpenetration, 4.1.1, all non-fully interpenetrated designs were considered to be confounded. Defining confounding (Definition 4.1.3) allows us to determine whether a survey design is fully confounded based on the random effects design matrix, $\mathbf{Z}$. Thus we can now determine directly whether it is possible to make an estimate of the interviewer effect. Given that we can now determine fully confounded designs we know that all other designs must be at least partially interpenetrated. The next section will define interpenetration in detail and discuss how to determine the degree of interpenetration associated with given designs.

4.1.5 Defining Interpenetration and Partial Interpenetration

Based on our conceptualization of both intergroup and intragroup interpenetration we can define full interpenetration as follows

Definition 4.1.4. *Full interviewer interpenetration occurs within spatial regions when all observations within each region are collected by different interviewers and interviewers collect data from more than one region.*

Definition 4.1.4 attempts to capture the concept that full interpenetration maximizes the number of elements v_{ij} of $\mathbf{V}$ in which the interviewer and spatial effects appear separately by requiring that there is no confounding in any element v_{ij} of $\mathbf{V}$. The extra condition that interviewers collect data from more than one spatial region is included so that we get repeated measurement of the individual interviewers and hence the interviewer effect will not be confounded with the residual variation.

After defining full interpenetration (Definition 4.1.4) and full confounding (Definition 4.1.2) all other scenarios must fall into the category of partial, or incomplete, interpenetration, which can therefore be defined as follows

Definition 4.1.5. *Interviewers are at least partially interpenetrated with spatial regions when more than one interviewer is allocated to at least one of the spatial regions and at least one interviewer collects data from multiple respondents in a single spatial region.*

We can now apply (4.4) to Definition 4.1.5 to determine interpenetration solely in terms of the the design matrix $\mathbf{Z}$.

Definition 4.1.6. *Let $\mathbf{Z_{wk}}$ refer to the columns of the design matrix, $\mathbf{Z}$, that relate to spatial zones and let $\mathbf{Z_{int}}$ refer to the columns of $\mathbf{Z}$ that relate to the interviewer. Then interviewers effects are at least partially interpenetrated with spatial effects if $\mathbf{Z_{int}Z_{int}^T} \neq \mathbf{Z_{wk}Z_{wk}^T}$.*

Note that Definition 4.1.6 follows directly from Definition 4.1.3 as all designs that are not fully confounded must be at least partially interpenetrated.

We have already seen that in the case of full interpenetration, we are aiming to maximize the number of elements of $\mathbf{V}$ in which the interviewer and spatial effects can be isolated by eliminating all confounding in non-diagonal elements of $\mathbf{V}$. Consequently under the GLMM the only way in which we can avoid observing both the interviewer effect and spatial effect simultaneously, in all possible elements of $\mathbf{V}$, is for all observations within a single spatial zone to be collected by different interviewers. Based on this we can then also express Definition 4.1.4 in terms of the design matrix, $\mathbf{Z}$.

Definition 4.1.7. *Let $\mathbf{Z_{wk}}$ refer to the columns of the design matrix, $\mathbf{Z}$, that relate to the spatial zones and let $\mathbf{Z_{int}}$ refer to the columns of $\mathbf{Z}$ that relate to the interviewer. Then interviewers are fully interpenetrated within spatial zones when all corresponding non-zero non-diagonal elements in $\mathbf{Z_{int}Z_{int}^T}$ and $\mathbf{Z_{wk}Z_{wk}^T}$ are not equal.*

By way of example consider $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ of differing dimension. If we had six observations divided into two spatial zones and allocated to three interviewers, we could represent this system as follows

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \\ Interviewer_3 \end{bmatrix}$$

Adopting Definition 4.1.4 we can see this is a scenario with full intergroup interpenetration but not full intragroup interpenetration because each spatial zone is enumerated by two different interviewers, rather than the three (a different interviewer for each within zone observation) that would have been required for full interpenetration. Now if we split $\mathbf{Z}$ into matrices detailing the columns of $\mathbf{Z}$ relating

to the spatial zones, $\mathbf{Z_{wk}}$, and the interviewers, $\mathbf{Z_{int}}$, i.e.

$$\mathbf{Z_{wk}}=\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{Z_{int}}=\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Then given $\mathbf{Z}$ must be constructed such that the row sums of both $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ must be equal to 1, this implies that the only way in which we can then construct a $\mathbf{Z}$ matrix containing non-interpenetrated groups is when $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ have columns in common. Consequently it follows from (4.3) that full confounding will only occur when all columns of $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ are equal, or if $\mathbf{Z_{wk}} = \mathbf{Z_{int}}$ when the columns of $\mathbf{Z_{int}}$ and $\mathbf{Z_{wk}}$ are sorted appropriately. Note that this condition may sometimes be more straightforward to evaluate than $\mathbf{Z_{int}}\mathbf{Z_{int}^T}=\mathbf{Z_{wk}}\mathbf{Z_{wk}^T}$ as was required in Definition 4.1.3 although it requires careful sorting of $\mathbf{Z_{int}}$ and $\mathbf{Z_{wk}}$ and may be confusing for $\mathbf{Z_{int}}$ and $\mathbf{Z_{wk}}$ of differing dimension.

In summary we have conceptualized interpenetration as a requirement for some form of repeated measurement to allow separable estimation of variance components such as the interviewer effect. In general standard definitions of interpenetration, such as Definition 4.1.1, have traditionally required hierarchical nesting of classification structures. Our new description of interpenetration encourages a more broad conceptualization that does not rely on nesting. The following section will briefly explore how hierarchical nesting relates to interpenetration and why it can be limiting with modern estimation techniques.

4.1.6 Interpenetration under Hierarchical Nesting

Under hierarchical nesting full interpenetration still occurs when the spatial effect does not appear together with the interviewer effect in any elements of $\mathbf{V}$ other than the diagonal elements. However hierarchical nesting also implies that each

interviewer belongs to only a single spatial zone or alternatively that each spatial zone is enumerated by only a single interviewer. Define hierarchical nesting as

Definition 4.1.8. *The classification structure implied by* $\mathbf{Z_{int}}$ *is hierarchically nested within the classification structure implied by* $\mathbf{Z_{wk}}$ *when there is more than 1 interviewer within at least one spatial zone or workload and interviewers in* $\mathbf{Z_{int}}$ *do not appear in more than one spatial zone as specified by* $\mathbf{Z_{wk}}$*.*

Then we can see that hierarchial nesting implies that all columns of $\mathbf{Z_{wk}}$ can be expressed as a linear combination of columns of $\mathbf{Z_{int}}$. Under hierarchical nesting of interviewers within spatial zones there would be a notional maximum of $\sum_{wk} n_{wk}\left(n_{wk}-1\right)$, elements of $\mathbf{V}$ (i.e. all non-diagonal elements for each spatial zone in a block diagonal variance covariance matrix) in which the interviewer effect and the spatial effect appear separately. However this would lead to confounding of the interviewer effect with the residual variation.

We can demonstrate this with a simple example in which we consider a hierarchically nested scenario with four observations and two spatial zones. Hierarchical nesting of interviewers within spatial zones implies that all interviewers only collect data from one spatial zone and as there are only two observations in each spatial zone in this case we need two different interviewers in each spatial zone to avoid confounding of the interviewer and spatial effects, giving us a total of four interviewers. The design matrix in this case would be

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \\ Interviewer_3 \\ Interviewer_4 \end{bmatrix}$$

which has an associated variance covariance matrix of

$$\mathbf{V} = \begin{bmatrix} \sigma^2_{wk} + \sigma^2_D & \sigma^2_{wk} & 0 & 0 \\ \sigma^2_{wk} & \sigma^2_{wk} + \sigma^2_D & 0 & 0 \\ 0 & 0 & \sigma^2_{wk} + \sigma^2_D & \sigma^2_{wk} \\ 0 & 0 & \sigma^2_{wk} & \sigma^2_{wk} + \sigma^2_D \end{bmatrix} \tag{4.10}$$

where $\sigma^2_D = \sigma^2_{int} + \sigma^2_\varepsilon$. We can see in (4.10) that we have 16 elements of $\mathbf{V}$ of which 8 are non-zero and can be used to estimate the variance components. In this hierarchically nested example we can estimate the spatial effect, σ^2_{wk}, from the $\sum_{wk} n_{wk}(n_{wk} - 1) = 4$ elements of $\mathbf{V}$ in which the spatial effect appears apart from the interviewer effect. Note that $\mathbf{V}$ is symmetric so these four elements are the same two elements repeated twice. In the remaining four non-zero elements of $\mathbf{V}$ the interviewer effect does not appear independently from the residual variation and hence we can only estimate the combined effect, σ^2_D rather than σ^2_{int} or σ^2_ε. This occurs because full interpenetration requires each observation, within a single spatial zone, to be collected by a different interviewer while hierarchical nesting requires that interviewers collect data from only one spatial zone. Consequently we can never achieve full intragroup interpenetration under hierarchical nesting without confounding the interviewer effect with the residual variation.

In comparison when we do not have hierarchical nesting of the classification structures it is possible for there to be more than $n_{wk}(n_{wk} - 1)$ observations which can be used in the estimation of either the interviewer or spatial effects. This scenario can occur under cross-classification with larger groups, for example in the cross-classified scenario below

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \\ Interviewer_3 \end{bmatrix}$$

which has an associated variance covariance matrix of

$$\mathbf{V} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_C^2 & \sigma_{wk}^2 & 0 & \sigma_{\text{int}}^2 \\ \sigma_{wk}^2 & \sigma_\varepsilon^2 + \sigma_C^2 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 + \sigma_C^2 & \sigma_{wk}^2 \\ \sigma_{\text{int}}^2 & 0 & \sigma_{wk}^2 & \sigma_\varepsilon^2 + \sigma_C^2 \end{bmatrix} \tag{4.11}$$

and we can see in (4.11) that we can estimate all of the variance components in this example as they are all isolated in some elements of $\mathbf{V}$. We now have 10 non-zero elements of $\mathbf{V}$ ($\sum_{wk} n_{wk} (n_{wk} - 1) = 4$, $\sum_{int} n_{int} (n_{int} - 1) = 2$ and $n = 4$ diagonal elements) from which estimates of the variance components can be made. However these 10 elements in the variance covariance matrix only relate to four observations in the original dataset, suggesting that any estimates made based on this data structure would be very unreliable. We can also see that (4.10) has full intragroup interpenetration while (4.11) does not, but our example without full intragroup interpenetration provides us with a larger number of observations in $\mathbf{V}$ from which to estimate the variance components. This implies that we cannot compare the degree of intragroup interpenetration to a by-zone baseline figure because the relationship between intragroup interpenetration and estimability of the interviewer effect is also dependent on the degree of intergroup interpenetration.

Intergroup interpenetration, on the other hand, occurs when the spatial effect appears separately from the interviewer effect in some but not necessarily all elements of each section of $\mathbf{V}$ relating to a particular spatial region.

Remark 1. The *section* of $\mathbf{V}$ influenced by a particular region or interviewer can be indicated by non-zero elements of $\mathbf{Z_i Z_i^T}$ where $\mathbf{Z_i}$ refers to the ith column of the design matrix $\mathbf{Z}$ and corresponds to the ith row of the random effect vector $\mathbf{u}$. This allows us to specify the contribution of a particular spatial region or interviewer to $\mathbf{V}$.

Thus a natural measure of the degree of intergroup interpenetration is to count the number of groups or regions that are at least partially interpenetrated. Hence we can define the degree of intergroup interpenetration as follows

Definition 4.1.9. *The degree of intergroup interviewer interpenetration within spatial zones is the proportion out of all spatial zones that contain more than one interviewer.*

Applying (4.4) to Definition 4.1.9 we can also express the degree of intergroup interpenetration in terms of the design matrix of the random effects, $\mathbf{Z}$

Definition 4.1.10. *The degree of intergroup interpenetration between two cross classified structures implied by* $\mathbf{Z_{int}}$ *and* $\mathbf{Z_{wk}}$ *will be equal to the proportion of all columns in* $\mathbf{Z_{int}}$ *that* $\mathbf{Z_{int}}$ *does not have in common with* $\mathbf{Z_{wk}}$*.*

We can therefore summarize these results as a single definition which allows us to assess the degree of intergroup interpenetration between the two cross classified structures based on the columns of the design matrix, $\mathbf{Z}$

Definition 4.1.11. *Given a Generalized Linear Mixed Model (2.12) in which the design matrix,* $\mathbf{Z}$ *has been partitioned into columns relating to random effects,* $\mathbf{Z_{int}}$ *and* $\mathbf{Z_{wk}}$*. There will be*

- *full confounding between the interviewers and the spatial zones if* $\mathbf{Z_{int}}\mathbf{Z_{int}^T} = \mathbf{Z_{wk}}\mathbf{Z_{wk}^T}$*.*
- *full interpenetration of interviewers and spatial zones if all corresponding non-zero non-diagonal elements in* $\mathbf{Z_{int}}\mathbf{Z_{int}^T}$ *and* $\mathbf{Z_{wk}}\mathbf{Z_{wk}^T}$ *are not equal.*
- *full intergroup interpenetration between interviewers and spatial zones if no columns of* $\mathbf{Z_{int}}$ *are equal to any columns of* $\mathbf{Z_{wk}}$*.*

The degree of intergroup interpenetration between two cross classified structures implied by $\mathbf{Z_{int}}$ *and* $\mathbf{Z_{wk}}$ *will be equal to the proportion of all columns in* $\mathbf{Z_{int}}$ *that* $\mathbf{Z_{int}}$ *does not have in common with* $\mathbf{Z_{wk}}$*.*

Thus depending on the design, it is possible that we may have more observations from which we can attempt to isolate the interviewer effect under partial rather than full interpenetration. Full interpenetration does not therefore imply the most appropriate estimate of the interviewer effect. Instead a combination of several

factors such as the survey design, the degree of intergroup interpenetration, the degree of intragroup interpenetration and the size of the variance components, will determine the accuracy of the interviewer effect estimate. The interaction of these design factors on the accuracy of interviewer effect estimates will be examined in more detail in Chapter 5.

The following section presents a number of tools that can be used to assess the degree of interpenetration for any given survey design and examples demonstrating how to calculate the degree of interpenetration based on these results.

4.1.7 Assessing the Degree of Interpenetration in Practice

This section will introduce techniques for determining the degree of interpenetration between two cross-classified structures. These concepts will then be applied to Definition 4.1.11.

As $\mathbf{Z_{int}}$ and $\mathbf{Z_{wk}}$ are generally matrices of differing dimension, we can consider Definition 4.1.11 via the impact of $\mathbf{Z_{int}}$ and $\mathbf{Z_{wk}}$ on the variance covariance matrix, $\mathbf{V}$. If we recall that the degree of interpenetration can be assessed by considering $\mathbf{V}$, then we can examine the element v_{ij} in $\mathbf{V}$, that relates to a particular variance component by considering the relevant columns of the design matrix, $\mathbf{Z}$. Then as non-zero elements of $\mathbf{Z_{int}Z_{int}^T}$ indicate contribution from the interviewer effect to a particular element of $\mathbf{V}$ (and similarly that non-zero elements of $\mathbf{Z_{wk}Z_{wk}^T}$ indicate contribution from the spatial effect to a particular element of $\mathbf{V}$) we can construct a matrix, $\mathbf{S}$, to consider the overlap, in terms of the contribution to elements of $\mathbf{V}$, between any two classification levels, such as the interviewers and the spatial zones, in a survey design

$$\mathbf{S} = abs\left(\mathbf{Z_{int}Z_{int}^T} - \mathbf{Z_{wk}Z_{wk}^T}\right)$$

where for any $n \times q$ matrix $\mathbf{M} = (\mathbf{m_{ij}})_{\mathbf{n\times q}}$, $abs\left(\mathbf{M}\right)$ is a matrix containing element by element absolute values for matrix $\mathbf{M}$, i.e. $abs(m_{ij})_{n\times q} = \left(|m_{ij}|\right)_{n\times q}$. Applying Definition 4.1.3 to the matrix, $\mathbf{S}$, we can see that all elements, s_{ij} in $\mathbf{S}$ will be zero if there is full confounding. This follows as Definition 4.1.3 implies $\mathbf{Z_{int}Z_{int}^T} = \mathbf{Z_{wk}Z_{wk}^T}$, and then $\mathbf{S} = abs(\mathbf{0}_{n\times q}) = \mathbf{0_{n\times q}}$. Similarly we can show by contradiction that some

elements of $\mathbf{S}$ will be non-zero when we have at least partial interpenetration. Our conceptualization of interpenetration then implies that the higher the number of non-zero elements in $\mathbf{S}$ the greater the degree of intragroup interpenetration and Definition 4.1.4 implies that there must be more than one non-zero element in each section (see Remark 1) of $\mathbf{S}$ relating to each spatial zone, if we are to have full intergroup interpenetration.

Thus in order to assess the degree of intergroup interpenetration within each spatial zone or workload of the spatial classification structure specified in $\mathbf{Z_{wk}}$ we can sum the elements of $\mathbf{S}$ over each spatial zone by forming the interpenetration matrix, $\mathbf{C_{wk}}$

$$\mathbf{C_{wk}}= \mathbf{Z_{wk}^T} abs\left(\mathbf{Z_{wk}}\mathbf{Z_{wk}^T}-\mathbf{Z_{int}}\mathbf{Z_{int}^T}\right)\mathbf{Z_{wk}}$$

Similarly the interpenetration matrix for the interviewers, or $\mathbf{C_{int}}$ can therefore be defined as

$$\mathbf{C_{int}}= \mathbf{Z_{int}^T} abs\left(\mathbf{Z_{wk}}\mathbf{Z_{wk}^T}-\mathbf{Z_{int}}\mathbf{Z_{int}^T}\right)\mathbf{Z_{int}}$$

then by construction we can see that the ith row of $\mathbf{C_{wk}}$ indicates the degree of interpenetration in the ith spatial zone or workload specified by the design matrix, $\mathbf{Z}$ by totalling the number of non-confounded observations in $\mathbf{V}$ corresponding to that spatial zone or workload. Expanding on this we can see that the element in the ith row and jth column of $\mathbf{C_{wk}}$ will be zero if there is no cross-classification between the interviewer and spatial zone classification structures implied by $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ that allows separable estimation of any observations in either spatial zone, $Zone_i$ or $Zone_j$ of the spatial classification structure indicated by $\mathbf{Z_{wk}}$. Non-zero elements in element (i, j) of $\mathbf{C_{wk}}$ indicate the presence of traditional hierarchical interpenetration (if any elements for which $j = i$ are non-zero) or the ability to isolate the variance components through cross-classification (if any elements for which $j \neq i$ are non-zero).

As an example, consider four individuals living in a single spatial zone (workload), who are enumerated by two interviewers. One possible representation of this

scenario would be with the following partially interpenetrated design matrix, $\mathbf{Z}$

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

The columns of $\mathbf{Z}$ that relate to the interviewer, $\mathbf{Z_{wk}}$, and the workloads $\mathbf{Z_{int}}$, are

$$\mathbf{Z_{wk}}= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \text{ and } \mathbf{Z_{int}}= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

and in this case $\mathbf{Z_{int}}$ has two rather than the four columns required for full interpenetration and hence from Definition 4.1.4 we only have partial interpenetration in this case. The contribution of each of the variance components to $\mathbf{V}$ is

$$\mathbf{Z}_{\text{wk}}\mathbf{Z}_{\text{wk}}^T= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{Z}_{\text{int}}\mathbf{Z}_{\text{int}}^T= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

which indicates that $\mathbf{S}$ is

$$\mathbf{S} = abs\left(\mathbf{Z}_{wk}\mathbf{Z}_{wk}^T - \mathbf{Z}_{\text{int}}\mathbf{Z}_{\text{int}}^T\right) = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

and the interpenetration matrices are

$$\mathbf{C_{wk}}= \mathbf{Z_{wk}^T} abs(\mathbf{Z_{wk}}\mathbf{Z_{wk}^T} - \mathbf{Z_{int}}\mathbf{Z_{int}^T})\mathbf{Z_{wk}} = 8$$

$$\mathbf{C_{int}}= \mathbf{Z_{int}^T} abs(\mathbf{Z_{wk}}\mathbf{Z_{wk}^T} - \mathbf{Z_{int}}\mathbf{Z_{int}^T})\mathbf{Z_{int}} = \begin{bmatrix} 0 & 4 \\ 4 & 0 \end{bmatrix}$$

In the above example we can see that $\mathbf{C_{wk}} = 8$, which means that there are 8 elements in $\mathbf{V}$ in which the interviewer and spatial effects are not confounded and

from which we can make an estimate of the interviewer and spatial effects. The way in which the interpenetration matrix $\mathbf{C_{wk}}$ can be interpreted is that each element c_{ij} in $\mathbf{C_{wk}}$ indicates the number of non-confounded elements in $\mathbf{V}$ that can be obtained through cross-classification of the spatial zones indicated by the ith and jth columns of $\mathbf{Z_{wk}}$. Hence we have can see that the 8 elements indicated by $\mathbf{C_{wk}}$ all belong to a single spatial zone and consequently we have hierarchical nesting according to Definition 4.1.8. We can also see that we have full intergroup interpenetration as all spatial regions are interpenetrated, although we do not have full intragroup interpenetration in this case. Note also that $\mathbf{V}$ is symmetric so although we have 8 elements of $\mathbf{V}$ in which the interviewer and spatial effects are not confounded, this only gives us four effective observations from which to estimate these effects.

If we consider $\mathbf{C_{int}}$ we can see that $c_{ij} = 0$ for the diagonal elements, $i = j$ and $c_{ij} = 4$ for non-diagonal elements, $i \neq j$. Summing all elements c_{ij} of $\mathbf{C_{int}}$ again indicates that we have a total of 8 elements in $\mathbf{V}$ in which the interviewer and spatial effects are not confounded and from which we can make an estimate of both the interviewer and spatial effects. Furthermore the first row (or column) of $\mathbf{C_{int}}$ indicates the level of interpenetration for the first interviewer's allocation and we can see that $c_{ij} = 4$ when $i \neq j$ indicating there is cross-classification of four elements of $\mathbf{V}$ allowing separation of the interviewer and spatial effects. We can also see that no columns of $\mathbf{C_{int}}$ are vectors with all elements equal to zero and hence we have full intergroup interpenetration for the interviewer classification structure.

4.1.8 Summary

The difference between the contribution of two variance components to $\mathbf{V}$ is $\mathbf{S}$. Any non-zero elements in $\mathbf{S}$ indicate one element in $\mathbf{V}$ in which the interviewer and spatial effects are not confounded and from which we can attempt to isolate these two variance components. We can summarize the number of distinct observations, with regards to the relevant variance components, in the variance covariance matrix, $\mathbf{V}$, by considering the interpenetration matrices $\mathbf{C_{wk}}$ and $\mathbf{C_{int}}$.

4.2 Determining the Degree of Intergroup Interpenetration

We have seen that the concept of partial interpenetration is an important one as we can now make estimates of the interviewer effect even when we have not designed for full interpenetration. We have introduced comprehensive definitions of both interpenetration and confounding and discussed both intergroup and intragroup interpenetration. The following section will now adopt these definitions and demonstrate how to estimate the degree of intergroup interpenetration based on the design matrix, $\mathbf{Z}$.

Given any two classification structures implied by $\mathbf{Z_{wk}}$ and $\mathbf{Z_{int}}$ containing n_{wk} and n_{int} groups respectively we can quickly determine the degree of intergroup interpenetration by calculating, $\mathbf{F_{wk}}$, the number of columns of the interpenetration matrix $\mathbf{C_{wk}}$ which contain at least one non-zero element and $\mathbf{F_{int}}$ the number of columns in $\mathbf{C_{int}}$ which contain at least one non-zero element. Then

- if $\mathbf{F_{wk}} = 0$ and $\mathbf{F_{int}} = 0$ the two classification structures are fully confounded and the variance components relating to these classification structures cannot be isolated,
- if $\mathbf{F_{wk}} = n_{wk}$ and $\mathbf{F_{int}} = n_{int}$ the two classification structures have full intergroup interpenetration and the accuracy of estimates of the variance components relating to these classification structures will depend on the degree of intragroup interpenetration, the size of the variance components, the survey design, the variability of the group sizes and the estimation procedure adopted,
- in all other cases the two classification structures have partial intergroup interpenetration, i.e. either $0 < \mathbf{F_{wk}} < n_{wk}$ or $0 < \mathbf{F_{int}} < n_{int}$. In this case the accuracy of estimates of the variance components relating to these classification structures will depend on the degree of both intergroup and intragroup interpenetration, the size of the variance components, the survey design, the variability of the group sizes and the estimation procedure adopted.

Since any confounded groups must be confounded for both the interviewer and the spatial classification structures , $n_{wk} - \mathbf{F_{wk}}$ will be equal to $n_{int} - \mathbf{F_{int}}$. Also to consider the degree of interpenetration as discussed in Definition 4.1.11 we will generally want to consider the proportion of interpenetrated groups, for example $d_Z = \frac{\mathbf{F_{wk}}}{n_{wk}}$ in the case of the spatial zones, rather than just the total number of interpenetrated groups, as it can be misleading to consider the number of interpenetrated groups without knowing the total number of groups.

4.2.1 Common Classification Structures

This section will demonstrate the calculation of $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$ in a number of different scenarios relating to common structures and levels of cross-classification;

Fully Confounded Scenario

The following two design matrices present a simple fully confounded scenario

$$\mathbf{Z_{int}} = \mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We can then calculate $\mathbf{S}$, $\mathbf{C_{int}}$, $\mathbf{C_{wk}}$, $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$ to demonstrate how these concepts can be used to confirm that these structures are fully confounded

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and therefore both interpenetration matrices will also have all elements equal to zero

$$\mathbf{C_{int}} = \mathbf{C_{wk}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus we can see that there are no rows or columns of either interpenetration matrix that contain any non-zero elements and consequently $\mathbf{F_{wk}} = \mathbf{F_{int}} = 0$ indicating, as expected, that the two structures are fully confounded.

In this case we used the interpenetration matrices to calculate the number of elements in the variance-covariance matrix in which the interviewer effect appears separately from the spatial effect based on the design matrix $\mathbf{Z}$. In the confounded scenario we have seen that there are no observations from which we can isolate the two random effects. In this simple scenario it is straightforward to assess the degree of confounding directly from the design matrix, $\mathbf{Z}$. However in more complex scenarios it may be more convenient to assess the degree of confounding via the interpenetration matrix.

The Hierarchical Linear Model (HLM)

The HLM is a commonly applied structure in which information regarding the nesting of two levels can be used to isolate the variance component estimates. We can demonstrate this situation with the example of the following two design matrices

$$\mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

and let $\mathbf{Z_{int}} = \mathbf{I_6}$. Then $\mathbf{S}$ is

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and we can see that the interpenetration matrix, $\mathbf{C_{int}}$, will be equal to $\mathbf{S}$ as $\mathbf{Z_{int}} = \mathbf{Z_{int}^T} = \mathbf{I_6}$, while $\mathbf{C_{wk}}$ is

$$\mathbf{C_{wk}} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

then in this case $\mathbf{F_{wk}} = 3$ and $\mathbf{F_{int}} = 6$ indicating that the two structures have full intergroup interpenetration, or in other words we have at least one repeated measurement within each spatial zone allowing us to separately estimate the effects.

Partially Interpenetrated Cross-Classified Structures

This final scenario is the one we are faced with when we attempt to estimate the interviewer effect on the CURF dataset. We have two cross-classified structures which have not been designed to be interpenetrated. However, some of the groups are interpenetrated to some degree. A simple example demonstrating this scenario

is presented in the following two matrices

$$\mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{Z_{int}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Now we can see in the design matrices above that the two spatial zones corresponding to the first and the third column of $\mathbf{Z_{wk}}$ are equivalent in design (i.e. the corresponding columns of the design matrix are equal) to two of the interviewer allocations in $\mathbf{Z_{int}}$, in this case indicated by the first and the fourth columns of $\mathbf{Z_{int}}$. These groups are not interpenetrated. In comparison the remaining spatial zone corresponding to the second column of $\mathbf{Z_{wk}}$ contains two distinct interviewers as indicated by the second and the third columns of $\mathbf{Z_{int}}$, meaning that we have repeated measurement of this spatial zone and hence interpenetration in this zone. Thus we can see that the classification structures are quite similar in this case and we would expect $\mathbf{S}$ to contain a high proportion of zero elements

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

In this case the interpenetration matrices are

$$\mathbf{C_{wk}}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \mathbf{C_{int}}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore $\mathbf{F_{wk}} = 1$ and $\mathbf{F_{int}} = 2$ which indicates that we do not have full intergroup interpenetration in this case. The degree of intragroup interpenetration, as indicated by the magnitude of the non-zero elements in $\mathbf{C_{wk}}$ and $\mathbf{C_{int}}$, is also quite low and any estimates of the interviewer effect will effectively be based on two (as $\mathbf{V}$ is symmetric) observations in this case.

4.2.2 Calculation of $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$ through Matrix Partitioning

For large datasets, such as the CURF, the $\mathbf{S}$ matrix created as an interim step during the estimation of $\mathbf{F_{int}}$ may be too large to efficiently calculate. For example in the case of the CURF, estimation of $\mathbf{S}$ would require storage of a (18,264 x 18,264) matrix. This may be beyond the storage capacity of many PCs and for larger datasets, such as the full LFS from which the CURF was derived, $\mathbf{S}$ will become even more difficult to consider. Consequently in this section computationally efficient techniques for calculating $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$ are introduced.

In practice it is clear that we do not need to explicitly calculate $\mathbf{S}$ to estimate $\mathbf{F_{int}}$, as it only requires storage of $\mathbf{C_{int}}$, generally a much smaller matrix. Hence we do not need to permanently store $\mathbf{S}$ and memory problems can be overcome using partitioning algorithms to independently store components such as the columns of $\mathbf{S}$.

The following steps were used to partition $\mathbf{S}$ into column vectors so that $\mathbf{F_{int}}$ could be calculated for the CURF dataset. Note that similar steps can be followed to calculate $\mathbf{F_{wk}}$.

1. Split $\mathbf{Z_{int}}\mathbf{Z_{int}}^{\mathbf{T}}$ and $\mathbf{Z_{wk}}\mathbf{Z_{wk}}^{\mathbf{T}}$ into $i = 1, ..., n$ column vectors by multiplying

$\mathbf{Z_{int}}$ with the ith column of $\mathbf{Z_{int}}^{\mathbf{T}}$, which we will denote by $\mathbf{Z_{int}}^{\mathbf{T}}{}_i$. i.e.

$$\mathbf{Z_{int}}\mathbf{Z_{int}}^{\mathbf{T}} = \left[\begin{array}{ccccc} \mathbf{Z_{int}}.\mathbf{Z_{int}}^{\mathbf{T}}{}_1 & \cdots & \mathbf{Z_{int}}.\mathbf{Z_{int}}^{\mathbf{T}}{}_i & \cdots & \mathbf{Z_{int}}.\mathbf{Z_{int}}^{\mathbf{T}}{}_n \end{array} \right]$$

Similarly split $\mathbf{Z_{wk}}\mathbf{Z_{wk}}^{\mathbf{T}}$ into $i = 1, ..., n$ column vectors by multiplying $\mathbf{Z_{wk}}$ with the ith column of $\mathbf{Z_{wk}}^{\mathbf{T}}$, again denoted by $\mathbf{Z_{wk}}^{\mathbf{T}}{}_i$.

2. Based on this expansion we can then simply express each column of $\mathbf{S}$, or $\mathbf{S}_i$, as

$$\mathbf{S}_i = abs\left(\mathbf{Z_{int}}.\mathbf{Z_{int}}^{\mathbf{T}}{}_i - \mathbf{Z_{wk}}.\mathbf{Z_{wk}}^{\mathbf{T}}{}_i\right)$$

3. Calculate $\mathbf{Z_{int}}^{\mathbf{T}}\mathbf{S}_i$ a (n_p x 1) vector (where p is the number of interviewers, i.e. $p = n_{int}$) equal to the ith column of $\mathbf{Z_{int}}^{\mathbf{T}} abs\left(\mathbf{Z_{int}}\mathbf{Z_{int}}^{\mathbf{T}} - \mathbf{Z_{wk}}\mathbf{Z_{wk}}^{\mathbf{T}}\right)$. Note that $\mathbf{Z_{int}}^{\mathbf{T}} abs\left(\mathbf{Z_{int}}\mathbf{Z_{int}}^{\mathbf{T}} - \mathbf{Z_{wk}}\mathbf{Z_{wk}}^{\mathbf{T}}\right)$ is a (n_p x n) matrix requiring a factor of $\frac{n}{n_p}$ less storage capacity than $\mathbf{S}$.

4. Based on $\mathbf{Z_{int}}^{\mathbf{T}} abs\left(\mathbf{Z_{int}}\mathbf{Z_{int}}^{\mathbf{T}} - \mathbf{Z_{wk}}\mathbf{Z_{wk}}^{\mathbf{T}}\right)$ we can directly calculate the interpenetration matrix $\mathbf{C_{int}} = \mathbf{Z}^{\mathbf{T}}_{\mathbf{int}} abs(\mathbf{Z_{wk}}\mathbf{Z}^{\mathbf{T}}_{\mathbf{wk}} - \mathbf{Z_{int}}\mathbf{Z}^{\mathbf{T}}_{\mathbf{int}})\mathbf{Z_{int}}$

5. Calculate $\mathbf{F_{int}}$ as the number of columns (or rows) of $\mathbf{C_{int}}$ which contain at least one non-zero element.

Note that more elaborate variance partitioning structures may be required to further reduce computer memory requirements for even larger datasets. These techniques can be further extended as required.

A copy of the algorithm used to calculate $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$ for the CURF can be found in Appendix B.

4.2.3 Degree of Interpenetration in the CURF

This section will present an example application of the concepts discussed in this chapter to determine the degree of interpenetration in the CURF dataset introduced in Chapter 3.

When we compare the interviewer and workload classification structures for the repeated panel data we get $\mathbf{F_{wk}} = 44$ which means that we can get repeated measurement (i.e. there is some form of interviewer interpenetration) in 44 out of the 50 workloads, which gives us a degree of intergroup interpenetration, d_Z for the workloads of 0.88. This agrees with what we saw in Table 3.3, indicating that there is a high degree of interviewer interpenetration in the workloads when we consider the repeated panel data.

Looking at the interviewer classification structures we get $\mathbf{F_{int}} = 103$ which means that we can get repeated measurement for 103 out of the 109 interviewers, which gives us a degree of intergroup interpenetration, d_Z for the interviewers of 0.95. Again this indicates a high degree of workload interpenetration in the interviewers when we consider the repeated panel data.

In comparison when we do not utilize the repeated panel information in the CURF and consider only the first month we get $\mathbf{F_{int}} = 8$ and $\mathbf{F_{wk}} = 9$ which means that we can get repeated measurement in 9 out of the 50 workloads (i.e. $d_Z = 0.18$) and for 8 out of the 49 interviewers ($d_Z = 0.16$). Thus we can clearly see that there is only a low degree of interpenetration between the workload and interviewer classification structures in a single month.

4.3 Discussion

We have produced a more rigorous definition of interpenetration based on the concept that confounding implies interviewer effects and spatial effects never appear separately and cannot be isolated. We then showed how the degree of interpenetration can be determined solely from the design matrix, $\mathbf{Z}$, in the GLMM (2.12). Based on the design matrix, we then constructed an interpenetration matrix, $\mathbf{C_{int}}$, to indicate the degree of interpenetration between any two classification structures. More concise summaries such as $\mathbf{F_{int}}$ or d_Z can also be constructed to indicate the degree of intergroup interpenetration. This information can be used by the survey methodologist to design more appropriate surveys or by the analyst to determine when the

interviewer effect can be isolated. For example when the degree of interpenetration can be determined, decisions regarding the appropriateness of interviewer effect estimates in partially interpenetrated surveys can then be explicitly considered. An application is presented in Chapter 5.

Although, $\mathbf{F_{int}}$, is a useful indicator of the level of interpenetration in a dataset it does not tell us the level of intragroup interpenetration, only the number of groups which contain some degree of interpenetration. This information is available in the interpenetration matrix, $\mathbf{C_{int}}$, but is difficult to summarize consistently. Hence we can explore the level of interpenetration by quoting d_Z. However to produce an appropriate estimate of the interviewer effect we must also have information about the effect of varying the degree of interpenetration on the variance of interviewer effect estimates.

We have seen in this chapter how to define partial interpenetration and how the level of interpenetration can be determined for a given survey design. As have also seen in Chapter 3 that designs with lower degrees of interpenetration appear to be associated with more uncertain (higher variance) estimates of the interviewer effect, the following chapter will now apply the definitions presented in this chapter to assess the effect of partial interpenetration on the estimation of the interviewer effect.

Chapter 5

Optimal Partially Interpenetrating Design

We have seen in Chapter 3 that we can apply modern estimation techniques to produce an interviewer effect estimate in partially interpenetrated surveys. We have also discussed how most surveys are not interpenetrated in practice due to cost constraints. Consequently it may be more cost effective to design partially interpenetrated surveys for the estimation of the interviewer effect.

We have now also comprehensively defined interpenetration and confounding for the first time and this will enable us to consider the impact of partially interpenetrated designs on estimates of the interviewer effect. This chapter will consider the impact of partially interpenetrated designs on the variance of interviewer effect estimates before optimal interpenetrated designs will be derived based upon a representative cost function.

5.1 The Impact of Partial Interpenetration

Under a moments approach (see for example Snijders and Bosker (1999)) we could consider the interviewer effect as the variance of the interviewer level residuals. Thus high levels of both intergroup and intragroup interpenetration are necessary as we are estimating each interviewer residual based on the number of intragroup observa-

tions and the interviewer effect based on the variation of the interviewer residuals. Consequently if we have low intergroup interpenetration our interviewer effect estimate will be based on few interviewer level estimates, effectively n_{int} observations, while if we have low intragroup interpenetration the individual interviewer level estimates may not be appropriate. Specification of this problem as a GLMM enables us to borrow strength from other groups, but as a general rule the higher the number of interpenetrated groups the better our estimate of the interviewer effect.

Under the GLMM (2.12) we have already seen that the variance covariance matrix, $\mathbf{V}$, can be partitioned into components representing

- The survey design. This is captured in the random effect design matrix, $\mathbf{Z}$.
- The magnitude of the variance components. This is captured in the variance component matrix, $\mathbf{D}$.
- The magnitude of the residual variation. This is captured in the residual matrix, $\mathbf{R}$.

so that $\mathbf{V} = \mathbf{ZDZ^T} + \mathbf{R}$ (see Chapter 2). Then given that the degree of interpenetration relates only to the design matrix, $\mathbf{Z}$, we can see that any change to the degree of interpenetration will also effect the variance covariance matrix, $\mathbf{V}$, from which estimates of $\mathbf{D}$ are subsequently isolated. Thus the degree of interpenetration influences our ability to estimate the interviewer effect no matter the magnitude of the interviewer effect.

In practice variance estimates associated with an interviewer effect estimate can be obtained from the information matrix based on the distribution of the response variable and given estimates of $\mathbf{D}$, $\mathbf{R}$ and $\mathbf{Z}$. We can therefore calculate a Variance Inflation Factor (vif) for estimates derived under a given survey design compared with the estimates we would have obtained under full interpenetration to assess the impact of interpenetration on the estimation of the interviewer effect. That is given a partially interpenetrated survey design and its associated design matrix, $\mathbf{Z}$, let $\mathbf{Z}^*$ be a design matrix describing the same spatial structure in which all observations within a single spatial zone are collected by different interviewers. Then $\mathbf{Z}^*$ will be

fully interpenetrated and comparable with the partially interpenetrated $\mathbf{Z}$. Then if $Var(\hat{\sigma}^2_{int})_{\mathbf{Z}}$ is the variance of the estimate of the interviewer effect under $\mathbf{Z}$ the vif can be calculated as

$$vif_{\mathbf{Z}} = \frac{Var(\hat{\sigma}^2_{int})_{\mathbf{Z}}}{Var(\hat{\sigma}^2_{int})_{\mathbf{Z}^*}} \tag{5.1}$$

McCullagh and Searle (2001) show that there is no general expression for the information of the random effects for all possible response distributions and hence variance estimates of the random effect estimates must be considered on a case by case scenario. As an example consider a normally distributed response variable, for which each element in the ith row and jth column of the information matrix for the random effect is (see Chapter 2 for more detail)

$$I(\sigma^2)_{\{i,j\}} = \frac{1}{2} tr(\mathbf{V^{-1}Z_iZ_i^TV^{-1}Z_jZ_j^T}) \tag{5.2}$$

Calculation of (5.2) requires an estimate of $\mathbf{V^{-1}}$, which in general requires information regarding both $\mathbf{D}$ and $\mathbf{R}$. In practice this means that we need to know the magnitude of all of the random effects in order to properly assess the effect of partial interpenetration on the variance of our interviewer effect estimates. In other words we need prior knowledge of the magnitude of the residual variation and the interviewer and spatial effects in order to appropriately design a survey to estimate the interviewer effect. Although in some cases prior knowledge may give us an approximate idea as to the magnitude of the interviewer effect, in general this information will not be available to the survey methodologist.

Thus we cannot compare vifs for the interviewer effect estimates for a number of competing survey designs before we actually conduct the survey. Hence in order to design optimal interpenetrating surveys for the purpose of estimating the interviewer effect we either need approximate prior estimates of the magnitude of all random effects or we need to be able to make general statements regarding the relationship between the degree of interpenetration and the vif associated with the interviewer effect estimates no matter the *true* magnitude of the random effects.

We have seen in Chapter 3 that estimates of $\mathbf{V}$ can be strongly biased when considering non-linear response variables. Consequently using the information matrix

to estimate the vif associated with an interviewer effect estimate for a given survey design may also be misleading. The following section will consider how appropriate empirical estimates of the vif associated with interviewer effect estimates can be made. These empirical techniques will be applied to derive the relationship between the degree of intergroup interpenetration and the vif associated with estimates of the interviewer effect for binary response variables. This relationship will then be applied to consider optimal partially interpenetrating designs for the estimation of the interviewer effect.

5.2 Optimal Design for Binary Response Multilevel Models

This section will consider current research into optimal design in multilevel models. Optimization is generally performed with respect to some stated property of the model, such as minimizing the variance of an estimator, conditional on a specified criteria, for example a representative cost function. In this chapter we will consider both single and multiple-objective optimal designs.

In their paper on the design of a multilevel study of communities Sastry *et al.* (2003, p 4) present a review of the literature on sample design in multilevel models, stating that it

'*... builds on standard sample survey methods (e.g. Cochran (1977); Kish (1965)) to consider several additional concerns relevant for estimating multilevel models.*'

but despite increasing application of multilevel modelling there has only been limited research into design issues relating to multilevel modelling. This was recognized by Moerbeek *et al.* (2001a, p 17) who state

'*Relatively little attention has so far been given to the planning of experiments in multilevel populations.*'

So far papers examining multilevel survey design have either been simulation based empirical studies, e.g. Mok (1995); Afshartous (1995); Normand and Zou (2002) or theoretical expositions considering the accuracy of the fixed effect parameter

estimates conditioning on cost and design, e.g. Snijders and Bosker (1993, 1999); Cohen (1998); Moerbeek *et al.* (2000, 2001a,b); Moerbeek and Wong (2002).

In the empirical studies of Mok (1995); Afshartous (1995), simulation and subsampling was utilized to explore optimal design choices for specific educational datasets. In comparison Cohen (1998) adopted a theoretical approach deriving expressions for the variance of parameter estimates for all parameters in a hierarchical, balanced two level linear mixed model in terms of sample design cost parameters. The approach of Cohen (1998, p 271) has the advantage that it attempts to engender

'*an understanding of how the sample should be apportioned as the different parameters vary.*'

However, Cohen's assumptions that the response variable is normally distributed and that the variance of the second level variance component can be considered through a design-free asymptotic expression, are limiting when considering interviewer effects. In particular we must consider the effect of the survey design if we wish to assess the impact of partial interpenetration on the accuracy of our interviewer effect estimates.

An extension of the Cohen (1998) paper is presented by Moerbeek *et al.* (2001a) who derive a linearization for the variance of the fixed effect parameter in multilevel models with logistic response. Moerbeek *et al.* (2001a, p 18) state that

'*... optimal designs cannot be derived analytically for PQL and numerical integration*'

and as Marginal Quasi-Likelihood (MQL) estimates are generally biased when considering non-linear response variables they present a general methodology for empirically determining the sampling variance of parameters in the multilevel logistic model. Moerbeek *et al.* (2001a) conclude that design decisions based upon biased MQL linearization of the variance of the fixed effect parameter estimates will generally be similar to those that would have been determined empirically using unbiased estimation methods. However, simulation techniques should be applied to explore the implications of design scenarios on variance component parameters such as the interviewer effect.

In this chapter we will explore the relationship between the survey design and

estimates of the interviewer effect. Our initial focus will be to minimize the variance of estimates of the interviewer effect conditional on a representative cost function. For binary data items, such as we are faced with in the CURF, this entails producing viable estimates of a vif comparing interviewer effect estimates under competing survey designs. The following section will present a general empirical methodology for exploring the impact of survey design on estimates of the interviewer effect and thereby establish a relationship between the degree of interpenetration and the variance that can be associated with interviewer effect estimates. We have already seen that there can be substantial bias in estimates of the interviewer effect derived under the MQL procedure and that numerical integration techniques are to be preferred when faced with a non-linear response variable. Consequently, extending the work of Moerbeek *et al.* (2001a), design scenarios for the optimal estimation of the interviewer effect will be assessed through MCMC simulation techniques.

5.2.1 Tools for Examining the Effect of Interpenetration on Interviewer Effect Estimates

We have seen that analytical approximations to the variance of the interviewer effect estimate can be derived under MQL estimation, but that these estimates may be strongly biased. On our CURF dataset, in Chapter 3, we saw that exact estimation techniques such as MCMC estimation are to be preferred for estimating variance components of non-linear response variables and consequently we must then also apply these techniques to examine how the estimates of the interviewer effect change with varying degrees of interpenetration when considering binary response variables.

We have also seen in Chapter 4 how to assess the degree of interpenetration in a given survey design and we can use this information to compare estimates of the interviewer effect in competing survey designs. This section will introduce a simulation algorithm that can be used to control the degree of interpenetration in random effect design matrices associated with the GLMM (2.12) and introduce empirical methods for estimating the variance of interviewer effect estimates under MCMC

estimation techniques for both linear and non-linear response variables. We will then apply these techniques to estimate variance inflation factors for interviewer effect estimates when we are faced with binary response variables and consider optimal partially interpenetrated designs for the estimation of the interviewer effect.

Controlling the Degree of Interpenetration

This section will introduce an algorithm to produce survey designs with a known degree of interpenetration. These survey designs can then be applied to either collect or simulate data from a design structure with a desired degree of interpenetration and will be utilized in the comparison of a number of partially interpenetrated designs.

The general form of the GLMM was introduced as Model (2.12), where $\eta = \mathbf{X}\beta + \mathbf{Zu}$ and $\mathbf{Z}$ is a $(n \times t)$ known matrix, of which all elements are either 0 or 1, that serves as the design matrix for the random effects as specified in the $(t \times 1)$ random effect vector, $\mathbf{u}$. Chapter 4 showed how the design matrix for the random effects, $\mathbf{Z}$, can be used to indicate the degree of interpenetration between interviewers and spatial regions. Given a constant degree of intragroup interpenetration across all spatial groups (i.e. workloads) we can use the information provided in Definition 4.1.11 to engineer a random effect design matrix, $\mathbf{Z}$, with a desired degree of intergroup interviewer interpenetration.

Definition 4.1.11 indicates that to increase the degree of intergroup interpenetration we need to increase the proportion of columns in $\mathbf{Z_{int}}$ that are not shared by $\mathbf{Z_{wk}}$. We can do this in a systematic way provided we have a constant known form of intragroup interpenetration. In the following example we apply full intragroup interpenetration in which each respondent within a single spatial region is enumerated by a different interviewer. To avoid confounding of the interviewer effect and the residual variation, interviewers are also required to interview respondents in as many different workloads as possible. Other forms of intragroup interpenetration can also be considered in a similar way, for example a minimal form of traditional interpenetration (see Definition 4.1.1) is indicated by randomly allocating all respondents in

a single spatial zone to one of two different interviewers.

In this chapter we consider a simple crossed 3 level scenario, i.e. there are only two random effects described in the random effect vector $\mathbf{u}$ in this case, the interviewer effect, σ^2_{int}, which is normally distributed and independent from the normally distributed spatial effect, σ^2_{wk}. Then assume that the residual variation, σ^2_{ε}, is normally distributed and independent of the interviewer and spatial effects and that we have a single constant fixed effect term in the model. Assume also that there are n observations in our dataset, n_{int} interviewers and n_{wk} spatial regions, such that $n > n_{int} \geq n_{wk}$ and that all workloads are of equal size, i.e. $\bar{n} = \frac{n}{n_{wk}}$ as the effect on the measurement variance is minimized for equal group sizes. Then we can say that the random effect design matrix $\mathbf{Z}$ will have n rows and $n_{int} + n_{wk}$ columns, $\mathbf{Z_{int}}$ will have n rows and n_{int} columns and $\mathbf{Z_{wk}}$ will have n rows and n_{wk} columns.

A simple algorithm (henceforth referred to as Algorithm 5.2.1) for simulating a design matrix for a balanced 3 level crossed GLMM model, given n, n_{int}, n_{wk}, a degree of inter-workload interpenetration, d_Z, and full intragroup interpenetration is therefore

1. Sort the observations in the dataset by spatial zone (workload) so that in the rows of $\mathbf{Z_{wk}}$ all observations in the same workload are stacked vertically and occur in consecutive rows. For example in the following $\mathbf{Z_{wk}}$ the observations in rows 1 and 4 are in the workload corresponding to the first column of $\mathbf{Z_{wk}}$ while observations 2 and 3 are in the workload corresponding to the second column of $\mathbf{Z_{wk}}$.

$$\mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Once $\mathbf{Z_{wk}}$ has been sorted all observations in the same workload should be

stacked vertically and occur in consecutive rows in $\mathbf{Z_{wk}}$.

$$\mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

2. Determine the number of workloads that are not interpenetrated, or have only one interviewer allocated to them and let this number be called n_{cf}. This can be done by calculating $n_{cf} = n_{wk}(1 - d_Z)$.

3. Then let the first n_{cf} columns of $\mathbf{Z_{int}}$ be equal to the first n_{cf} columns of $\mathbf{Z_{wk}}$. The remaining $n_{int} - n_{cf}$ columns $\mathbf{Z_{int}}$ can then be constructed as follows. The first $n_{cf}\frac{n}{n_{wk}}$ rows of the remaining $n_{int} - n_{cf}$ columns $\mathbf{Z_{int}}$ contain elements that are all equal to 0. The remaining $n - n_{cf}\frac{n}{n_{wk}}$ rows can be constructed by stacking $x = \frac{n(1-\frac{n_{cf}}{n_{wk}})}{n_{\text{int}}-n_{cf}}$ copies of the identity matrix $\mathbf{I_{n_{int}-n_{cf}}}$. Note if $x(n_{int} - n_{cf}) > n - n_{cf}\frac{n}{n_{wk}}$ simply choose the first $n - n_{cf}\frac{n}{n_{wk}}$ rows. Note also that x measures the number of repeated measurements of each interviewer, so if x is small this will have an impact on the estimability of the interviewer effect.

An example of this algorithm follows; if we have $n = 4$ observations split into $n_{wk} = 2$ workloads, with $n_{int} = 3$ interviewers and given a degree of inter-workload interpenetration of $d_Z = 0.5$ we could start with the following sorted, balanced design matrix for the workloads

$$\mathbf{Z_{wk}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Then $n_{cf} = n_{wk}(1 - d_Z) = 2(1 - 0.5) = 1$ and hence the first column of $\mathbf{Z_{int}}$ is equal to the first column of $\mathbf{Z_{wk}}$. The first $n_{cf}\frac{n}{n_{wk}} = 1 * \frac{4}{2} = 2$ rows of the remaining $n_{int} - n_{cf} = 3 - 1 = 2$ columns of $\mathbf{Z_{int}}$ are comprised solely of elements

equal to 0, while the remaining $n - n_{cf}\frac{n}{n_{wk}} = 4 - 1 * \frac{4}{2} = 2$ rows are comprised of $x = \frac{n(1-\frac{n_{cf}}{n_{wk}})}{n_{\text{int}}-n_{cf}} = \frac{4(1-\frac{1}{2})}{3-1} = 1$ copy of the identity matrix $\mathbf{I_2}$. In other words

$$\mathbf{Z}_{\text{int}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus our random effect design matrix in this case would be

$$\mathbf{Zu} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Zone_1 \\ Zone_2 \\ Interviewer_1 \\ Interviewer_2 \\ Interviewer_3 \end{bmatrix}$$

We can now apply Definition 4.1.11 to the design matrix above to show that this design matrix has a degree of inter-workload interpenetration of $d_Z = 0.5$ and a degree of inter-interviewer interpenetration of $d_Z = 0.3$. This simple example would not be a suitable design for the estimation of the interviewer effect as it flouts the condition that $n >> n_{int} \geq n_{wk}$ and hence x is small and we would not get high levels of repeated measurement of interviewers in this case. However, this simple example has demonstrated how this algorithm can be applied to generate design matrices with a known degree of intergroup interpenetration.

Now that we have shown how to generate designs with a known degree of interpenetration in the simple balanced 3 level GLMM with crossed interviewer and spatial hierarchies we can use this algorithm to construct designs with controlled degrees of intergroup interpenetration. The following section will consider how we can determine the vifs to compare the estimation of the interviewer effect under these designs.

Determining the Variance of Interviewer Effect Estimates in Logistic Multilevel Models

We have seen that in order to compare the variance of the interviewer effect estimate under competing survey designs with binary response variables we must apply empirical techniques to determine the variance. The following technique is an adaptation of one presented by Moerbeek *et al.* (2001a).

Let $\hat{\sigma}^2_{int,r}$ be the rth estimate of the interviewer effect, σ^2_{int}, made under r independent simulations of the dataset with constant sample design and model parameters, then an empirical estimate of the variance can be produced by

$$\hat{\text{var}}(\hat{\sigma}^2_{int}) = \frac{1}{r-1}\sum\nolimits_r \left(\hat{\sigma}^2_{int,r} - \frac{\sum_r \hat{\sigma}^2_{int,r}}{r}\right)^2 \tag{5.3}$$

Note that similar empirical estimators can be constructed for the variance of the spatial effect and fixed effect estimates and for any estimation procedure. Now if we consider a simple 3 level Linear Mixed Model

$$y_{ijk} = \mu + \theta_j + \phi_k + \varepsilon_{ijk} \tag{5.4}$$

where, i, j, and k and indices referring to the classification levels in the data and y is a normally distributed response variable with independent and normally distributed random effects $\theta_j \sim N(0, \sigma^2_{\text{int}})$, $\phi_k \sim N(0, \sigma^2_{wk})$ and $\varepsilon_{ijk} \sim N(0, \sigma^2_{\varepsilon})$. Then Model (5.4) can be written in matrix form $\mathbf{y} = \beta\mathbf{X} + \mathbf{Zu} + \varepsilon$ with variance covariance matrix $\mathbf{V} = \mathbf{Z^T D Z} + \mathbf{R}$ and we can determine the variance of the interviewer effect estimate via the inverse of the information matrix. Here we adopt the notation of McCulloch and Searle (2001) in which $\{_m a_{ij}\}^r_{i,j=1}$ indicates a matrix with r rows, i, and r columns, j, comprising elements a_{ij}. E.g. $\{_m a_{ij}\}^2_{j=1} = [\ a_{i1} \quad a_{i2}\ ]$.

$$\mathrm{I}\begin{bmatrix} \hat{\beta} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{X^T V^{-1} X} & 0 \\ 0 & \frac{1}{2}\left\{_m tr\left[\mathbf{Z_i^T V^{-1} Z_j}\left(\mathbf{Z_i^T V^{-1} Z_j}\right)^{\mathbf{T}}\right]\right\}^r_{i,j=1} \end{bmatrix}$$

We can then compare the empirical variance estimates of the interviewer effect with theoretical estimates determined from the inverse of the information matrix in a simple example. If we simulate a fully interpenetrated design matrix containing 2450

observations partitioned into equal size groups of 70 interviewers and 50 workloads (i.e. set $d_Z = 1$, $n = 2450$, $n_{int} = 70$ and $n_{wk} = 50$ under Algorithm 5.2.1) and then simulate a normally distributed response variable according to Model (5.4) with a fixed effect size of 0.5, a individual level variance of 1, a workload (spatial) effect of 25 and an interviewer effect of 6.25, then the inverse of the information matrix will be

$$\mathrm{I}\begin{bmatrix} \hat{\beta} \\ \hat{\sigma}^2_{\mathrm{int}} \\ \hat{\sigma}^2_{wk} \end{bmatrix}^{-1} = \begin{bmatrix} 0.59 & 0 & 0 \\ 0 & 25.18 & -0.01 \\ 0 & -0.01 & 1.14 \end{bmatrix}$$

In comparison 500 independent simulations of the data lead to the following empirical variance estimates.

Parameter	Asymptotic Variance	Empirical Variance
$\hat{\beta}$	0.59	0.59
$\hat{\sigma}^2_{wk}$	25.18	26.07
$\hat{\sigma}^2_{int}$	1.14	1.20

Table 5.1: Comparison of Empirical Variance and Asymptotic Variance of Parameter Estimates in Fully Interpenetrated 3 Level Linear Mixed Model

MCMC estimation techniques were used to produce variance component estimates for each simulation in Table 5.1. These techniques have been applied exclusively in this chapter. We can see that in this example the empirical variance provides a good estimate of the true asymptotic variance of the parameter estimates. The empirical variance will generally be an overestimate, approaching the true variance as the number of simulations, r, increases. Although simulation may be computationally intensive we must also note that to calculate the asymptotic variance we have had to invert the variance covariance matrix $\mathbf{V}$, a $(n \times n)$ matrix, which, for large datasets, may be beyond the computational ability of many computers.

We have seen in this section how to estimate the variance of the interviewer effect estimate empirically. This is especially important for non-linear response variables in which the estimates of $\mathbf{V^{-1}}$ and therefore the asymptotic variance estimates for

the interviewer effect are generally biased. In practice many datasets, such as the CURF, contain non-linear variables and the following section will demonstrate how this empirical technique can be extended to estimate the variance of the interviewer effect for binary response variables.

Empirical Variance Estimates in Logistic Response Multilevel Models

We have demonstrated how we can empirically determine the variance of variance component estimates in multilevel models with normally distributed response variables. We will now apply these same techniques to the logistic response multilevel model and determine an appropriate simulation length, r, for estimation of the variance of interviewer effect estimates.

Consider the logistic representation of the balanced 3 level GLMM with crossed interviewer and spatial hierarchies

$$\Pr(y_{ijk} = 1|\pi_{jk}) = \frac{\exp(\pi_{jk})}{1 + \exp(\pi_{jk})} \tag{5.5}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the dataset.

$$\pi_{jk} = \mu + \phi_k + \theta_j$$

and

- μ is a fixed effect
- The random effects are independent and normally distributed, i.e. $\phi_k \sim N(0, \sigma^2_{wk})$ and $\theta_j \sim N(0, \sigma^2_{int})$
- i, j and k are indices referring to the person/individual level, the interviewer and workload levels respectively

We can again empirically estimate the variance of the variance component estimates by applying (5.3). Estimates of the variance of the interviewer effect estimates for a simple 3 level logistic response dataset simulated according to (5.5) with 5000 observations, 100 interviewers and 50 workloads (i.e. the design matrix is determined

by setting $d_Z = 1$, $n = 5000$, $n_{int} = 100$ and $n_{wk} = 50$ under Algorithm 5.2.1), a fixed effect of 3 and random effects of $\sigma^2_{int} = 1$ and $\sigma^2_{wk} = 4$ can be seen in Table 5.2 for various simulation lengths, r. For each simulation MCMC estimation was used to produce and estimate of σ^2_{int}.

Number of Simulations r	Empirical variance
400	0.047
300	0.047
200	0.047
100	0.054
50	0.065
25	0.057
10	0.072

Table 5.2: Empirical Variance of Interviewer Effect Estimates in Multilevel Logistic Response Model

We can see in Table 5.2 that the estimated variance of the interviewer effect estimate stabilizes as we increase the number of simulations. The choice of the number of simulations will therefore be a compromise between the time required to run further simulations, r, and the desired accuracy of the variance estimate. So under this scenario a sensible number of simulations would be at least $r = 200$, however any changes to the underlying population parameters, or the design, would require re-evaluation of this simulation length.

Variance Inflation Factors For Logistic Response

We have demonstrated how to estimate the variance of the interviewer effect estimate under the logistic response multilevel model and we will now use these techniques to consider how the variance of the interviewer effect estimate is related to the degree of intergroup interpenetration.

A simple 3 level logistic response multilevel model was simulated, according to Model (5.5) and with a design matrix based on a known degree of inter-workload interpenetration as determined by Algorithm 5.2.1. The following parameter settings were applied, $\beta = 2.5$, $\sigma^2_{int} = 0.5^2$, $\sigma^2_{wk} = 1.5^2$, $n = 5000$, $n_{int} = 100$ and $n_{wk} = 50$

and full intragroup interpenetration was required for all interpenetrated groups. The empirical variance of the interviewer effect estimate based on a number of different degrees of inter-workload interpenetration (as determined by Algorithm 5.2.1) was then calculated via (5.3) using $r = 250$ simulations for each estimate. The variance inflation factors for the interviewer effect estimate against the degree of intergroup interpenetration can be seen in Figure 5.1 following

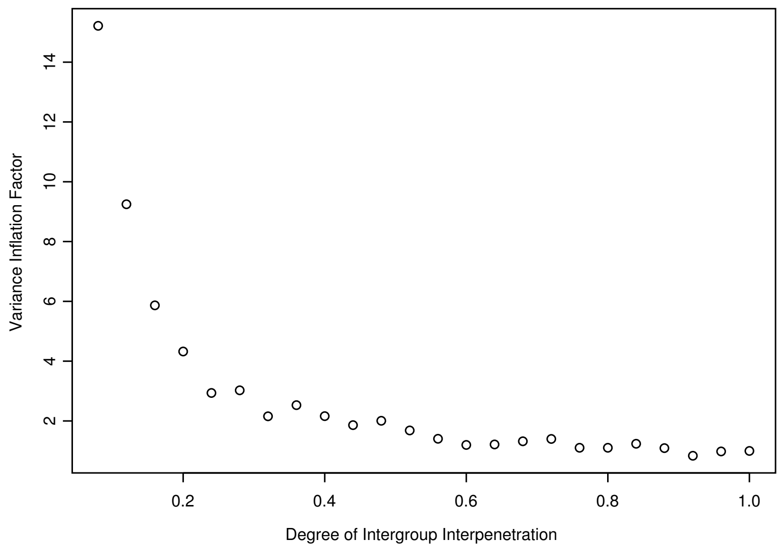

Figure 5.1: Variance Inflation Factor for Interviewer Effect Estimate by Degree of Intergroup Interpenetration: Logistic Response Model

In Figure 5.1 the variance inflation factor on the interviewer effect estimate increases as the degree of intergroup interpenetration decreases. However the variance inflation factor does not exceed 1.5 until less than 50% of the workloads are interpenetrated, i.e. the degree of intergroup interpenetration falls below 0.5. Moreover we can see that the variance inflation factor increases rapidly for lower degrees of intergroup interpenetration, asymptoting to infinity when there is full confounding. Consequently, under this simple model, we would potentially be able to make a reasonable estimate of the interviewer effect without full interpenetration.

Summary

Based on our definition of partial interpenetration we demonstrated in Chapter 4 how the level of interpenetration can be determined for any survey design. In this chapter we introduced a simple empirical framework for producing unbiased estimates of the vifs associated with interviewer effect estimates made under competing survey designs. Based on this we can now compare any two competing survey designs and through this determine the most appropriate survey design for the estimation of the interviewer effect.

As increased degrees of interpenetration are generally associated with increased travel costs this suggests a cost optimal partially interpenetrated design. The next section will examine the general relationship between the vif that can be associated with interviewer effect estimates, the total costs required to implement the survey and the sample size and use this relationship to consider optimal partially interpenetrated survey designs.

5.3 Optimal Design Based on Travel Cost Functions

We have seen in Figure 5.1 that there appears to be a relationship between the degree of intergroup interpenetration and the variance of the interviewer effect estimate. Given that increasing the degree of interpenetration associated with a survey is generally a costly exercise this section will consider simple cost functions that can be associated with a survey design. We will then consider how to minimize the variance of the interviewer effect estimate based on these cost functions and hence prepare optimal interpenetrating designs.

In the previous section we presented techniques that can be applied to compare survey designs for the estimation of the interviewer effect. Direct application of these techniques for optimal design purposes would require an estimate of the magnitude of the interviewer effect. However in practice we generally cannot ob-

tain this estimate until after the survey has been conducted. We therefore need to establish a general relationship that can be applied in practice to determine an optimal interpenetrating design for the estimation of the interviewer effect without prior knowledge of the magnitude of the interviewer effect. In this section we will explore the general relationship between the variance of interviewer effect estimates, the sample size and the degree of intergroup interpenetration in order to produce an approximate relationship that can be applied in practice to produce an optimal partially interpenetrated survey design based on a cost function.

5.3.1 Representative Travel Cost Function

Although full interpenetration will generally lead to a more reliable estimate of the interviewer effect (i.e. the estimate with the lowest variance inflation factor) this will also generally entail higher operational costs. This is because interpenetration requires interviewers to travel between workloads to interview respondents, leading to increased travel costs. A fully confounded design will generally minimize travel costs as all of the respondents enumerated by a single interviewer will reside in closer proximity to one another.

We will initially consider a simple representative cost function that models only the travel costs associated with a survey design, before considering more complex cost functions in the following sections. By way of example, assume the cost of an interviewer travelling to an interview in a different workload is four times that of the interviewer travelling to meet any new respondent within the same workload, i.e. let $c_1 = 1$ cost unit and $c_2 = 4$ cost units. This is a strong simplification, roughly equivalent to assuming all workloads are adjacent and that travel costs are constant. The total travel costs are related to how many workloads an interviewer travels to and how many observations the interviewer collects in each workload rather than just the total number of workloads. Thus if we assume that the travel costs associated with each interviewer's first interview are $c_2 = 4$ units then the total travel costs can be calculated as

$$C = \sum_i \left(c_1(a_i - b_i) + c_2 b_i\right) \tag{5.6}$$

where for each interviewer, i,

- c_1 is the cost of travelling to different respondents in the same workload,
- c_2 is cost of travelling between workloads,
- a_i is total number of respondents enumerated by a single interviewer, i and
- b_i is total number of workloads in which interviewer i conducts interviews. Consequently $\sum_i b_i \geq n_{wk}$ and reflects the form of interpenetration in the design. For example $\sum_i b_i = 2n_{wk}$ if two interviewers are randomly allocated to every workload while $\sum_i b_i = n_{wk}$ under full confounding.

If we assume the same design parameters as in Figure 5.1, i.e. $n = 5000$, $n_{int} = 100$, $n_{wk} = 50$ and full intragroup interpenetration for all interpenetrated groups (see Definition 4.1.4), then the design matrix, $\mathbf{Z}$, for any degree of inter-workload interpenetration, d_Z, can be determined by algorithm (5.2.1). The total number of respondents enumerated by each interviewer, a_i, and the total number of workloads each interviewer collects data from, b_i, are indicated in $\mathbf{Z}$ and the total travel costs associated with this design by (5.6). The relationship between the total travel costs required by the survey design and the degree of intergroup interpenetration, d_Z, can be seen in Figure 5.2 following.

Figure 5.2 shows that the total travel costs increase steadily as the degree of intergroup interpenetration also rises. This is not a simple linear relationship, as both the number of interviewers travelling between workloads as well as the number of workloads the interviewers travel to increase as the degree of intergroup interpenetration is increased. For example with two interviewers travelling to two interpenetrated regions $\sum_i b_i = 4$, while for three interviewers all travelling to three interpenetrated regions $\sum_i b_i = 9$. In practice $\sum_i b_i$ is generally straightforward to determine as information regarding the number of workloads interviewers travel to is provided in the design matrix, $\mathbf{Z}$.

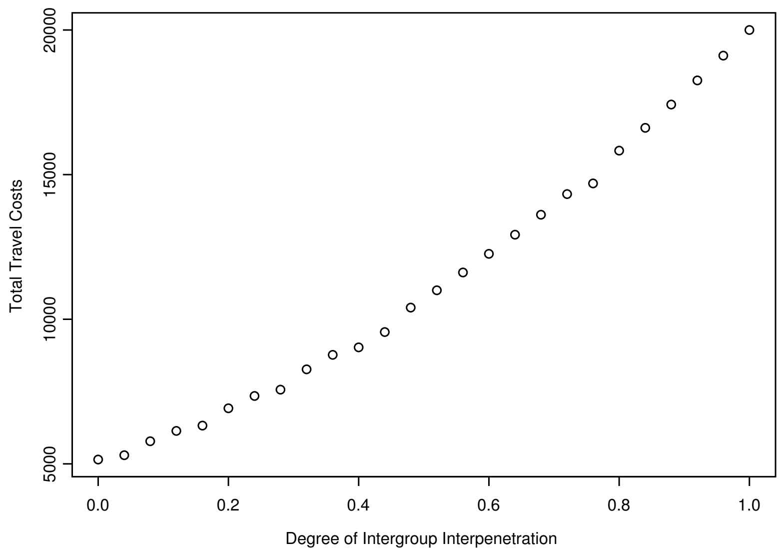

Figure 5.2: Total Travel Costs Implied by Degree of Intergroup Interpenetration

Summary

Figures 5.2 and 5.1 demonstrate that increasing the degree of intergroup interpenetration, while holding the degree of intragroup interpenetration in any interpenetrated groups constant, will lead to more reliable estimates of the interviewer effect, but also increase the costs associated with conducting the survey. Hence to produce interviewer effect estimates of a specified reliability it may be more cost effective to conduct a larger partially interpenetrated survey that allows us to collect data from more respondents rather than a smaller survey of similar cost that is fully interpenetrated. The optimal design choice will ultimately depend on properties of the underlying population, the aims of the survey designer and the total budget available.

5.3.2 Extended Cost Function Incorporating Sample Size

We have already seen for a fixed sample size, n, that increased levels of interpenetration will lead to higher costs while enabling us to more accurately estimate the interviewer effect. When faced with a fixed budget the methodologist can therefore

either choose to design a survey with an increased total sample size or with a higher degree of interpenetration. If the aim is to produce as accurate estimates of the interviewer effect as possible based on a given budget, this will usually imply an optimal degree of interpenetration. We can demonstrate this with a simple extension to our earlier cost function

$$C = (c_1 + c_3)\, n + (c_2 - c_1) \sum\nolimits_i b_i \tag{5.7}$$

where for each interviewer, i,

- C is the total cost,
- c_1 is the total cost associated with enumerating different respondents in the same workload,
- c_2 is cost of travelling to interview the first respondent in a different workload,
- c_3 is the total cost associated with including an extra respondent in the sample,
- b_i is total number of workloads in which interviewer i conducts interviews.

$\sum_i b_i$ is related to the degree of intergroup interpenetration, d_Z, implied by the random effects design matrix, $\mathbf{Z}$. If we assume our fixed body of interviewers is greater than the given number of workloads, i.e. $n_{int} > n_{wk}$, and that we have as full intragroup interpenetration as possible (this may be limited by the total number of interviewers that are available) in any interpenetrated groups then we can say that

$$\begin{aligned} \sum\nolimits_i b_i &= n_{wk} + n_{wk} d_Z \left(n_{int} - n_{wk} - 1 + n_{wk} d_Z\right) \\ &= n_{wk} \left\{1 + d_Z \left(n_{int} - 1 + n_{wk} \left[d_Z - 1\right]\right)\right\} \end{aligned} \tag{5.8}$$

This result recognizes that there must be at least one interviewer allocated to each of the n_{wk} regions. In the $n_{wk} d_z$ interpenetrated regions full intragroup interpenetration is assumed so that all remaining available interviewers (i.e. take the total

number of available interviewers minus the number who are already enumerating non-interpenetrated regions; $n_{int} - n_{wk} + n_{wk}d_z$) collect the data. Result (5.8) allows us to restrict consideration to only designs containing full intragroup interpenetration. Other forms of intragroup interpenetration can also be specified in a similar way, for example random allocation of two interviewers to each workload would lead to $\sum_i b_i = n_{wk}(1 + d_Z)$ provided $n_{int} \geq n_{wk}$ and full confounding can be represented as $\sum_i b_i = n_{wk}$ when $n_{int} = n_{wk}$.

In large scale surveys the number of workloads is generally determined geographically and can therefore be considered as fixed. The hiring and training of interviewers is a slow and costly process and hence for design purposes we will also consider the body of available interviewers to be fixed. Note that with longer lead-in periods it will be possible to prepare further interviewers, however this scenario remains to be considered at a later date. Given cost coefficient estimates, c_1, c_2 and c_3 the total cost is a simple function of both the degree of intergroup interpenetration and the sample size. Combining (5.7) and (5.8) then gives

$$C = (c_1 + c_3)\, n + (c_2 - c_1)\, n_{wk} \left\{1 + d_Z \left(n_{int} - 1 + n_{wk} \left[d_Z - 1\right]\right)\right\} \qquad (5.9)$$

Then as we have already seen seen in Figure 5.1 that the variance of the interviewer effect estimate is also a function of the degree of intergroup interpenetration, we can minimize this variance subject to the cost constraint (5.9) to determine the optimal degree of interpenetration for the estimation of the interviewer effect. To do this we first need to establish the relationship between the vif, which we are trying to minimize, and the remaining variables in the cost function (5.9). As c_1, c_2, c_3, n_{wk} and n_{int} are generally all fixed this means we need to establish the relationship between the variance inflation factor associated with the interviewer effect estimate, the sample size, n and the degree of intergroup interpenetration, d_Z. The following section will begin by examining the relationship between the variance of the interviewer effect estimate and the degree of intergroup interpenetration.

5.3.3 Relationship Between vif and d_Z

We have seen in Figure 5.2 that for given design parameters, n, n_{int}, n_{wk} and a form of intragroup interpenetration, the total costs associated with a survey are related to the degree of intergroup interpenetration. This section will establish a general relationship between the vif that can be associated with an interviewer effect estimate and the degree of interpenetration, d_Z.

Under the moments approach we could consider an estimate of the interviewer effect as the variance of the interviewer level mean estimates minus the contribution from other sources of variation in the model. See for example Snijders and Bosker (1999) for the moments estimators in the 2 level HLM. Consequently the more interviewers we observe the lower the variance that will be associated with the interviewer effect estimate. We also know that the higher the degree of interpenetration, the more individual estimates of interviewers we will be able to make. Thus as the degree of interpenetration increases we would expect the variance of the interviewer effect estimate to fall and there would be an approximate inverse relationship between vif and d_Z.

Figure 5.3 following shows the estimated relationship between the inverse of the variance inflation factor and the degree of intergroup interpenetration and can be used to estimate the approximate relationship. This data was simulated based on the simple three level logistic response multilevel model (5.5) and with a design matrix based on a known degree of inter-workload interpenetration as determined by Algorithm 5.2.1. The following parameter settings were applied, $\beta = 2.5$, $\sigma^2_{int} = 0.5^2$, $\sigma^2_{wk} = 1.5^2$, $n = 5000$, $n_{int} = 100$ while $n_{wk} = 50$ and full intragroup interpenetration was required for all interpenetrated groups.

Figure 5.3 suggests that the relationship between the inverse of the variance inflation factor and the degree of intergroup interpenetration is approximately linear. The fitted OLS regression line, based on the assumption that the residual is normally distributed, has an R-Squared of 0.9388 with an estimated intercept that is not significantly different from 0 and a slope that is not significantly different from 1.

A similar result was also determined for a normal response model. There is an

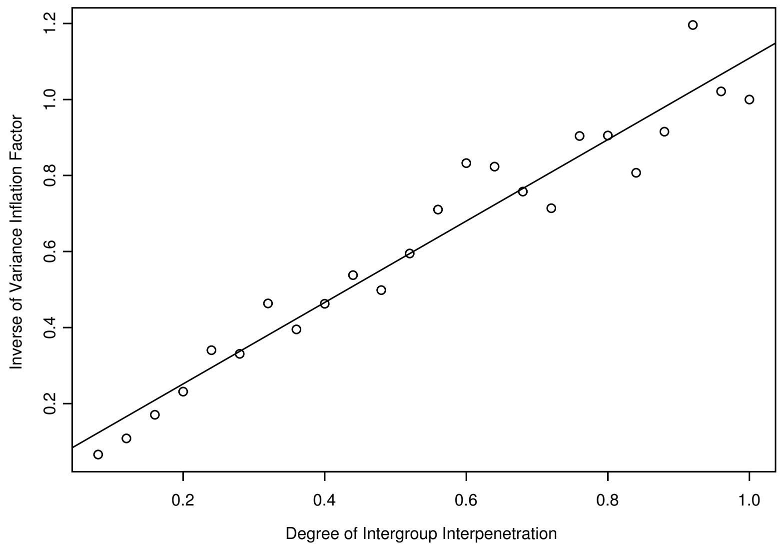

Figure 5.3: Inverse of Variance Inflation Factor for Interviewer Effect Estimate by Degree of Intergroup Interpenetration: Logistic Response Model

	Value	Std. Error	t value	P value
Intercept	0.0375	0.0354	1.0602	0.3006
Slope	1.0714	0.0583	18.3773	0.0000

Table 5.3: OLS Estimates: Inverse of Variance Inflation Factor as the Dependent Variable and Degree of Intergroup Interpenetration as the Explanatory Variable

inverse relationship between the variance inflation factor, vif, on the interviewer effect estimate and the degree of intergroup interpenetration, d_Z, conditional on a fixed degree of intragroup interpenetration in any interpenetrated groups. Thus if v_{dZ} is the variance of the interviewer effect estimate associated with a specific degree of intergroup interpenetration, d_Z, and, under our assumptions, also associated with a number of fixed design parameters, C, c_1, c_2, c_3, n, n_{int}, n_{wk} and the degree of intragroup interpenetration, we can see that

$$vif_{dZ} = \frac{v_{dZ}}{v_{dZ|dZ=1}} \simeq \frac{1}{d_Z} \tag{5.10}$$

Based on this approximate general relationship between the variance of the estimate of the interviewer effect and the degree of intergroup interpenetration we can now

make some statements regarding optimal design of partially interpenetrated surveys for the estimation of the interviewer effect. Recall, however, that larger sample sizes will generally lead to more accurate estimates of the interviewer effect under full interpenetration. In other words $v_{dZ|dZ=1}$ is also a function of the sample size, n, conditional on the population parameters, σ^2_{int}, σ^2_{wk}, σ^2_ε and μ, the fixed design parameters, C, c_1, c_2, c_3, n_i, n_{wk} and the degree of intragroup interpenetration as specified by $\sum_i b_i$. The following section will explore the relationship between the variance of the interviewer effect estimate under full interpenetration and the sample size and assess the implications of this relationship for optimal interpenetrating survey designs.

5.3.4 Relationship Between $v_{dZ|dZ=1}$ and the Sample Size

We can develop an initial analytic understanding as to the relationship between the variance of the interviewer effect estimate under full interpenetration and the sample size, by considering the simple case of a balanced 2 level (respondent at level 1 and interviewer at level 2) HLM for a normally distributed response variable. This can be done by adapting an asymptotic expression for the second level variance component. For example, if we take Longford's expression for the asymptotic expression for the second level variance component in the 2 level HLM (Longford (1993) p 58) and recognize that balanced designs imply all interviewers collect data from $\bar{n}$ respondents so that $\bar{n} = \frac{n}{n_{int}}$ we can write the asymptotic variance of the second level variance component as

$$\begin{aligned}
\mathrm{var}(\sigma^2_{\mathrm{int}}) &= \frac{2\left(\sigma^2_\varepsilon\right)^2}{n}\left(\frac{1}{\bar{n}-1} + 2\omega + \bar{n}\omega^2\right) \\
&= \frac{2\left(\sigma^2_\varepsilon\right)^2 n_{int}}{n\left(n-n_{int}\right)} + \frac{4\left(\sigma^2_\varepsilon\right)^2\omega}{n} + \frac{2\left(\sigma^2_\varepsilon\omega\right)^2}{n_{int}} \qquad (5.11)
\end{aligned}$$

Here $\omega = \frac{\sigma^2_{int}}{\sigma^2_\varepsilon}$ is a constant determined by the given population parameters σ^2_ε and σ^2_{int}. Given also that by definition $n \geq n_{int} \geq 1$ and holding constant n_{int} (as recruiting and training interviewers is generally much more costly than altering the sample size) we can then see in (5.11) that as the sample size, n, increases the

variance of the interviewer effect will approach $\frac{2(\sigma_\varepsilon^2\omega)^2}{n_{int}}$. However, generally we are worried about the impact of interviewer effects when cost constraints limit both the number of interviewers and the sample size such that $n > n_{\text{int}} > 1$ but the sample size is still not large. Furthermore, we saw in Chapter 3 that for the case of the CURF $\sigma_\varepsilon^2 > 1 > \sigma_{\text{int}}^2$. In general we cannot say anything about the actual magnitude of σ_ε^2 and σ_{int}^2, although we would generally expect that $\sigma_\varepsilon^2 > \sigma_{\text{int}}^2$ for carefully run surveys. In this case we would expect that (5.11) will be dominated by the $\frac{4(\sigma_\varepsilon^2)^2\omega}{n}$ term within the range of interest for the methodologist. Consequently we would expect that for the purposes of designing cost effective partially interpenetrated surveys there will be an approximately inverse relationship between $v_{dZ|dZ=1}$ and the sample size.

When we consider the logistic response multilevel model (5.5) a similar approximate relationship appears to hold. In this model the relationship between the sample size and the variance of the interviewer effect estimate is not immediately clear as the expansion under MQL is biased and as Moerbeek *et al.* (2001a) point out expressions for the variance of the interviewer effect cannot be derived analytically for PQL and numerical integration. The relationship between the variance of the interviewer effect estimate and the sample size can be determined empirically and this relationship is presented in Figure 5.4 following. In Figure 5.4 the data is simulated according to Model (5.5) with the following parameter settings $\mu = 2.5$, $\sigma_{int}^2 = 0.5^2$, $\sigma_{wk}^2 = 1.5^2$, $n_{int} = 100$, $n_{wk} = 50$, $d_Z = 1$ while full intragroup interpenetration was required for all interpenetrated groups. Note that for each simulation the design matrix was determined using algorithm (5.2.1).

We can see in Figure 5.4 that the variance of the interviewer effect estimate decreases as the sample size increases, eventually asymptoting to a level determined by the given design and population parameters. There appears to be an approximately inverse relationship between the variance of the interviewer effect estimate and the sample size and this relationship is presented in Figure 5.5 following

Figure 5.5 indicates that the relationship between the logarithm of the empirical variance of the interviewer effect estimate and the logarithm of the inverse of

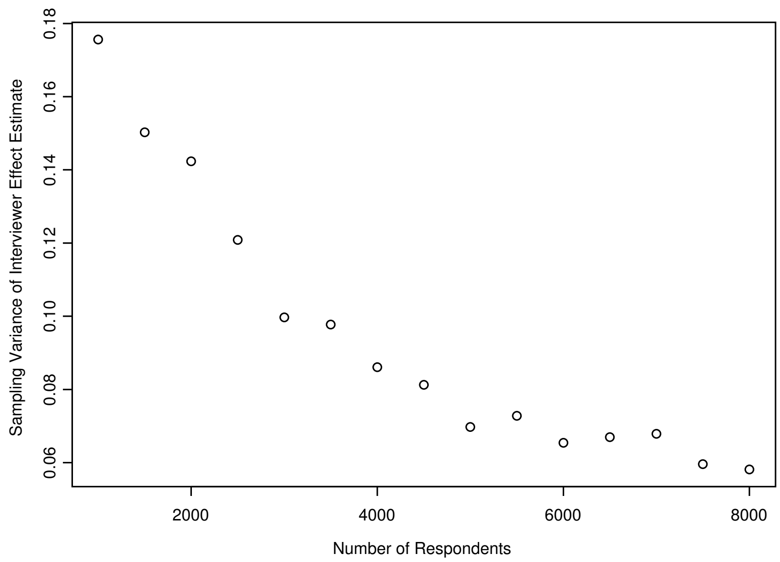

Figure 5.4: Variance of Interviewer Effect Estimate by Sample Size: Logistic Response Model under Full Interpenetration

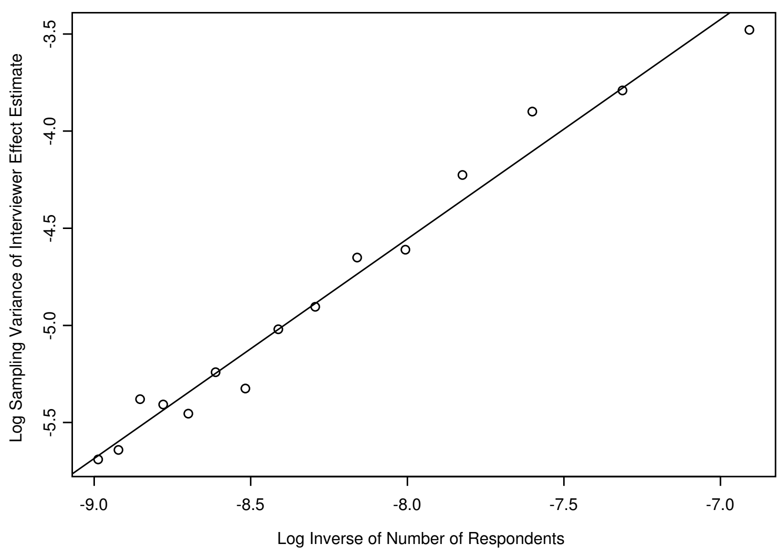

Figure 5.5: Log of Variance of Interviewer Effect Estimate by Log of Inverse of Sample Size: Logistic Response Model under Full Interpenetration

the total sample size is approximately linear and provides further support for an inverse relationship. The fitted OLS regression line, based on the assumption that

the residual is normally distributed, has an R-Squared of 0.977 and the estimated coefficients can be found in Table 5.4.

	Value	Std. Error	t value	P value
Intercept	4.4868	0.3983	11.2648	0.0000
Slope	1.1102	0.0481	23.4981	0.0000

Table 5.4: OLS Estimates: Log of Empirical Variance as the Dependent Variable and Log of the Inverse of Total Sample Size as the Explanatory Variable

There is an approximately inverse relationship between the variance of the interviewer effect estimate under full interpenetration and the total sample size. Thus

$$v_{dZ|dZ=1} \simeq \frac{a_z}{n} \tag{5.12}$$

where a_Z is a constant for any given set of population parameters, σ^2_{int}, σ^2_{wk}, σ^2_{ε}, μ, the fixed design parameters, n_{int}, n_{wk} and the degree of intragroup interpenetration as specified by $\sum_i b_i$. So in this case based on our parameter settings of $\mu = 2.5$, $\sigma^2_{int} = 0.5^2$, $\sigma^2_{wk} = 1.5^2$, $n_{int} = 100$, $n_{wk} = 50$, $d_Z = 1$ and assuming full interpenetration for each design determined by Algorithm 5.2.1, we can see in Table 5.4 that a_Z will be approximately equal to exp(4.4868) or 88.8.

We have seen in this section that there is a general approximately inverse relationship between the variance of interviewer effect estimates under full interpenetration and the sample size. We have also seen that the vif associated with an estimate of the interviewer effect is related to the degree of intergroup interpenetration, d_Z. Given that the vif has been defined as the ratio of the variance of the interviewer effect under a given degree of interpenetration to the variance of the interviewer effect under full interpenetration, we can now combine these relationships to express the relationship between the variance of an interviewer effect estimate, the degree of intergroup interpenetration and the sample size. The following section will explore this relationship in more detail.

5.3.5 Composite Relationship Between v_{dZ}, n and d_Z

By combining (5.12) and (5.10) we can now express an approximate relationship between the variance of the interviewer effect estimate in terms of the degree of intergroup interpenetration and the sample size. We can then use this information to minimize our cost function (5.9) with respect to the sample size and the degree of intergroup interpenetration. In this way we will be able to determine, for a given total budget, the optimal degree of interpenetration for producing estimates of the interviewer effect with the lowest possible variance.

Putting together (5.12) and (5.10) we get the following relationship between the variance of the interviewer effect estimate, the degree of intergroup interpenetration and the sample size

$$v_{dZ} \simeq \frac{a_z}{n.d_Z} \tag{5.13}$$

Then combining (5.13) and (5.9) we can express the variance of the interviewer effect estimate for a fixed total cost C in terms of the degree of intergroup interpenetration

$$v_{dZ} \simeq \frac{a_z(c_1 + c_3)}{d_z \left\{C - (c_2 - c_1)[n_{wk} + n_{wk}d_z(n_{int} - n_{wk} - 1 + n_{wk}d_z)]\right\}} \tag{5.14}$$

Note that for a fixed total cost the variance of the interviewer effect estimate can also be expressed as a function of the sample size, as (5.13) implies that

$$n \simeq \frac{C - (c_2 - c_1)[n_{wk} + n_{wk}d_z(n_{int} - n_{wk} - 1 + n_{wk}d_z)]}{(c_1 + c_3)} \tag{5.15}$$

To find the optimal degree of interpenetration we minimize the variance of the interviewer effect as expressed in Equation (5.14), for a given total budget, C. As the numerator of (5.14) is a constant this is equivalent to maximizing the denominator of (5.14) over the entire range of possible degrees of intergroup interpenetration, i.e. $0 < d_Z \leq 1$. As the denominator is a cubic expression in d_Z, its derivative will be quadratic and the maximum value within this range will occur either when we have full interpenetration, i.e. $d_Z = 1$ or at a local maximum which can be determined by one of the quadratic roots in Equation (5.16) following

$$\frac{(c_1 - c_2)n_{wk}[n_{int} - n_{wk} - 1] \pm \Delta_C}{3(c_2 - c_1)n_{wk}^2} \tag{5.16}$$

where $\Delta_C = \sqrt{((c_1 - c_2)n_{wk}[n_i - n_{wk} - 1])^2 - 3((c_2 - c_1)n_{wk}^2)((c_2 - c_1)n_{wk} - C)}$.

By way of example consider a binary response variable simulated according to Model (5.5) as if it was collected via a survey with a total budget of $C = 10000$ and design parameters $n_{int} = 100$, $n_{wk} = 50$, $c_1 = 1$, $c_2 = 4$, $c_3 = 2$, design matrices determined by algorithm (5.2.1), full intragroup interpenetration as specified by the form of $\sum_i b_i$ in (5.8) and finally $a_Z = 88.8$ (see Table 5.4). A plot of the variance of the interviewer effect estimate can be seen in Figure 5.6 following

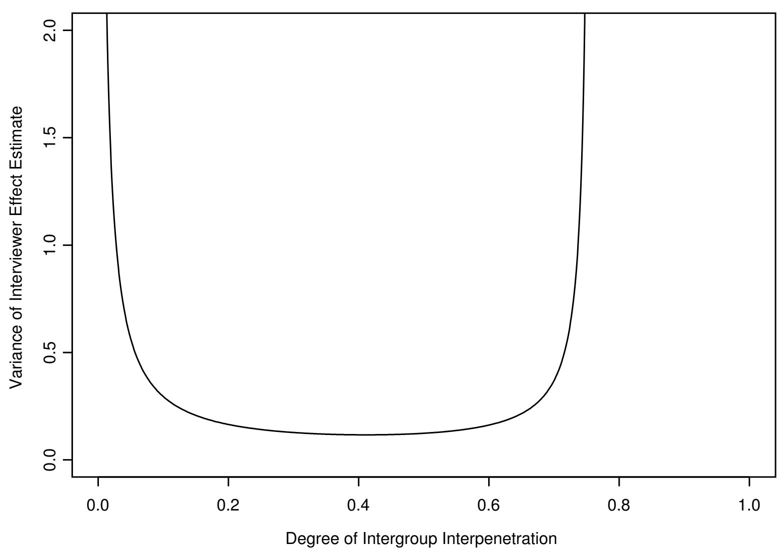

Figure 5.6: Variance of Interviewer Effect Estimate by Degree of Intergroup Interpenetration: Total Fixed Cost 10000

From Figure 5.6 we see that the total budget is not high enough for full interpenetration to be considered in this case as the highest degree of intergroup interpenetration affordable under this budget is $d_Z = 0.756$. Note also that as we approach full confounding, i.e. $d_Z \rightarrow 0$ it becomes harder to estimate the interviewer effect and the variance of the estimate approaches infinity. In comparison as we increase the degree of interpenetration we are forced to reduce our sample size, n, accordingly due to our total budget constraint. Consequently, as $d_Z \rightarrow 0.756$, $n \rightarrow 0$ and the impact of this small sample size is that the variance of the interviewer effect estimate again approaches infinity.

What is happening in Figure 5.6 is that when the sample size is high, for a fixed budget we can only afford a low degree of interpenetration, leading to an unreliable estimate of the interviewer effect. When the degree of interpenetration is high, however, we can only afford a small sample size, which again leads to unreliable estimates of the interviewer effect. Consequently we can see in Figure 5.6 that there is an optimal degree of intergroup interpenetration associated with the minimum possible variance of the estimate of the interviewer effect estimate given our total budget constraint. In this case the optimal degree of interpenetration is not at the end points of the range $0 < d_Z \leq 1$ and the optimal degree of interpenetration is at a local minimum which can be determined by Equation (5.16). This corresponds to a degree of intergroup interpenetration of $d_z = 0.411$ with an implied sample size that can be determined by Equation (5.15) i.e. $n = 1854$ and an optimal variance of the interviewer effect estimate of $v(\hat{\sigma}^2_{int}) = 0.12$. We can also see in Figure 5.6 that the variance of the interviewer effect does not increase rapidly as we move away from the optimal degree of interpenetration and hence degrees of interpenetration near the optimal may still be applied to produce reliable estimates of the interviewer effect.

We can utilize Equation (5.16) to determine the total minimum budget required before the optimal degree of interpenetration will be determined to occur when the survey is fully interpenetrated. In other words to estimate the interviewer effect with unlimited finances we would still require a total budget of at least

$$C \geq c_{12}n_{wk} + \frac{(3n^2_{wk}c_{12} + c_{12}n_{wk}[n_{int} - n_{wk} - 1])^2 - (c_{12}n_{wk}[n_{int} - n_{wk} - 1])^2}{3n^2_{wk}c_{12}} \tag{5.17}$$

where $c_{12} = c_2 - c_1$, for full interpenetration to be optimal. This reflects the point at which the positive root of (5.16) becomes greater than one. In the case of our example this means that full interpenetration is optimal under our cost function, design and population parameters when the total budget is greater than $C = 42,152$.

5.3.6 Effect of Optimal Interpenetrating Design on Sampling Variance of Mean

We have already seen that there is a cost associated with increasing the degree of interpenetration when designing a survey and hence under a fixed budget this will lead to a reduced sample size. In isolation we can use this information to determine an optimal degree of interpenetration for the estimation of the interviewer effect. In practice, however, any reduction in the sample size will have an impact on other components of the total survey error such as the sampling variance.

The sampling variance of the mean is usually quoted as the measure of variance associated with estimates derived from a survey. This section will examine the effect of optimal interpenetrating designs on both estimates of the sampling variance and the variance of the interviewer effect and discuss the cost implications for the survey designer.

Consider the Sampling Variance (SV) component of the Total Variance (TV) associated with the sample mean, $\bar{y}_s$. Under a Simple Random Sampling WithOut Replacement (SRSWOR) sampling scheme, the sampling variance will be

$$SV\left(\bar{y}_s\right) = \left(1 - \frac{n}{N}\right)\frac{S_Y^2}{n}$$

where we set the population size to be $N = 1000000$, n is the sample size and S_Y^2 is the adjusted population variance and therefore constant for a given population. Then given a fixed total budget C and information regarding the fixed design parameters, c_1, c_2, c_3, n_{int}, n_{wk} and the degree of intragroup interpenetration as specified by $\sum_i b_i$ we can calculate the optimal degree of intergroup interpenetration for estimating the interviewer effect under our fixed budget C. We can see from the cost function (5.7) that d_z determines n for a fixed C and consequently the degree of interpenetration determines the sample size and hence also impacts on the magnitude of the sampling variance.

We can then work out the variance inflation factors against what would be achieved at the optimal level of interpenetration for both the variance of the interviewer effect and the sampling variance. Table 5.5 following compares the variance

inflation factors for estimates of both the sampling variance and the variance of the interviewer effect, compared with the variance at the optimal degree of interpenetration for a binary response variable simulated according to Model (5.5), as if it was collected via a survey with a total cost of $C = 10000$ and design parameters $n_{int} = 100$, $n_{wk} = 50$, $c_1 = 1$, $c_2 = 4$, $c_3 = 2$, design matrices determined by algorithm (5.2.1), full intragroup interpenetration as specified by the form of $\sum_i b_i$ in (5.8) and finally $a_Z = 88.8$ (see Table 5.4).

d_Z	Sample Size	vif Variance of Interviewer Effect	vif Sampling Variance
0	3283	NA	0.563
0.1	3013	2.53	0.615
0.2	2693	1.41	0.688
0.3	2323	1.09	0.798
0.411	1854	1	1
0.5	1433	1.06	1.294
0.6	913	1.39	2.032
0.7	343	3.17	5.408
0.756	2	439.5	810.0

Table 5.5: Variance Inflation Factors Against Optimal Degree of Interpenetration under SRSWOR for Total Cost 10000

For our total budget of $C = 10000$ the optimal degree of intergroup interpenetration in this example is $d_Z = 0.411$ and the maximum degree of intergroup interpenetration affordable is $d_Z = 0.756$. We can see in Table 5.5 that as the sample size, n, increases the sampling variance decreases. Consequently we get a lower sampling variance component of the total survey error if we decrease the degree of intergroup interpenetration for a fixed budget constraint. However, if we decrease the degree of intergroup interpenetration past the optimal level, in this case $d_Z = 0.411$ then the reduction in sampling variance comes at the expense of reduced accuracy of the interviewer effect estimate. This relationship can be seen more clearly in Figure 5.7 following.

In Figure 5.7 the vif for the variance of the interviewer effect and the sampling variance are equal at a vif of 1 and a degree of intergroup interpenetration of $d_Z = 0.411$. This occurs because we have calculated the vifs with respect to the

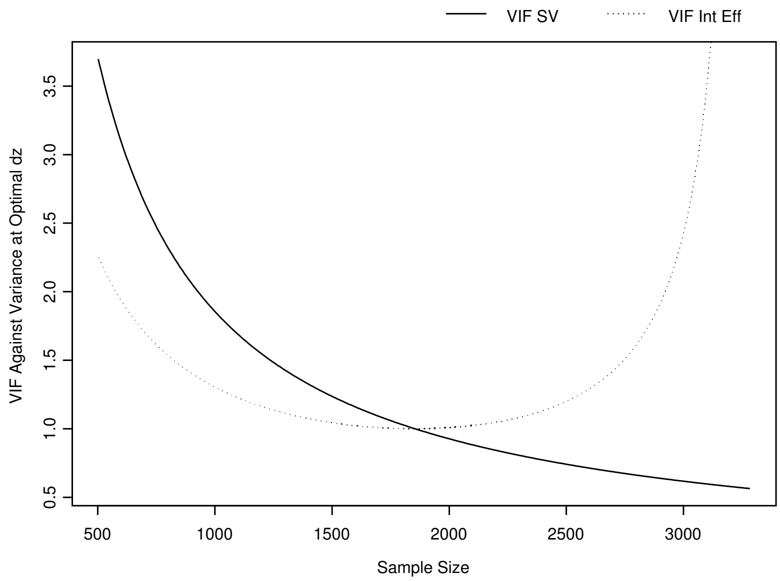

Figure 5.7: vifs Against Optimal Degree of Interpenetration for Sampling Variance and Variance of Interviewer Effect under SRSWOR for Total Cost 10000

optimal degree of interpenetration.

When the sample size in Figure 5.7 is maximized this corresponds to a degree of interpenetration of $d_z = 0$, as for a fixed budget we can only afford to increase the degree of interpenetration by reducing the sample size. Thus when $n = 3283$ we cannot afford any interpenetrated regions and we cannot produce an estimate of the interviewer effect. We can see in Figure 5.7 that this will lead to the lowest possible sampling variance. On the other hand if we reduced the sample size by a small margin, such as 270, we could afford a degree of interpenetration of $d_z = 0.1$, with an associated 9.2% increase in the sampling variance. With $d_z = 0.1$ the interviewer effect would be estimable, though with a vif of 2.53, so for a minor increase in the sampling variance we can greatly improve the reliability of estimates of the interviewer effect. If we wanted to produce the same estimate of the interviewer effect without altering the sampling variance (i.e. hold the sample size fixed) we can apply Equation (5.14) to show this could also be achieved by increasing the total budget by 9.6%

We can see in Figure 5.7 that as we move to the left of the optimal degree

of intergroup interpenetration, the variance inflation factor for both the sampling variance and the variance of the interviewer effect increase, indicating these points are sub-optimal for minimizing either the sampling variance or the variance of the interviewer effect estimate. Consequently if we are interested in both the sampling variance and the variance of the interviewer effect estimate, we would never design a survey with a degree of intergroup interpenetration higher than the single-objective optimal (i.e. the optimal degree of interpenetration chosen for the sole purpose of estimating the interviewer effect). On the other hand we can see that points to the right of the single-objective optimal degree of intergroup interpenetration lead to an increased variance for the interviewer effect estimate, but a decreased sampling variance. This implies that a degree of intergroup interpenetration less than the single-objective optimal may be preferred by the survey designer as it will lead to a lower sampling variance, even though the interviewer effect estimate will be less accurate than could have been achieved with the optimal degree of intergroup interpenetration. This will depend on the priorities of the survey designer as, for a fixed budget, they may choose to reduce the sampling variance at the expense of reduced accuracy of the interviewer effect estimate. From our example above we can see that when faced with a fixed budget of $C = 10000$ we could achieve a 20% reduction in the sampling variance by accepting a 10% increase in the variance of the interviewer effect estimate compared with the single-objective optimal position.

Multiple objective designs can also be prepared which aim to simultaneously minimize both the total variance of the mean (TV) and the variance of the interviewer effect estimate. However vifs associated with the TV depend on the relative magnitude of the sampling variance and the interviewer effect as $TV(\bar{y}_s) = \left(1 - \frac{n}{N}\right) \frac{S_Y^2}{n} + \frac{\sigma_\varepsilon^2}{n} + \frac{\sigma_{\text{int}}^2}{n_{\text{int}}}$. Thus, although the sampling variance and individual level measurement error reduce as the sample size increases, the magnitude of the interviewer effect term in the total variance, $\frac{\sigma_{\text{int}}^2}{n_{\text{int}}}$, is fixed for a constant body of interviewers. Consequently, for a fixed sample size, to reduce the contribution of the interviewer effect to the total variance we would need to increase the number of interviewers, n_{int}, collecting data in the survey. For a fixed body of interviewers

this suggests that Figure 5.7 presents a conservative relationship as the vif plot for the TV will be flatter than the vif plot for the SV. Hence it would generally be expected that varying the degree of interpenetration, and therefore the sample size, will have less of an effect on the TV. We must therefore make assumptions regarding the relative magnitude of the interviewer effect to prepare optimal multiple objective designs which minimize both the total variance and the variance of interviewer effect estimates.

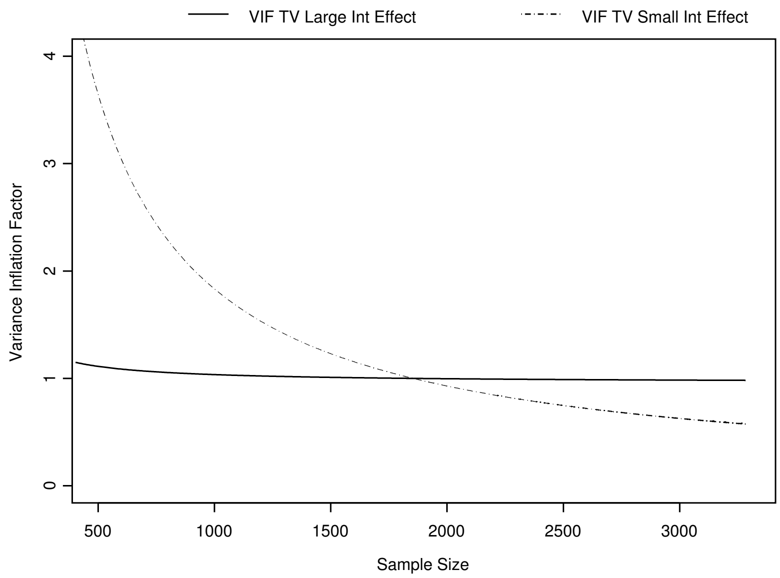

Figure 5.8: vifs for Total Variance with Interviewer Effect as Large or Small Proportion of Total Variance

In Figure 5.8 vifs are presented for the total variance against the sample size when the interviewer effect is both a high and a low proportion of the total variance. When the interviewer effect comprises the majority of the total variance, then increasing the sample size, without altering the number of interviewers conducting the survey, does not greatly affect the total variance. In this case it is important that the interviewer effect is estimated as it will be the major source of uncertainty in the survey. On the other hand, if the interviewer effect is a relatively small component of the total variance then the sampling variance dominates and increasing the sample size has a strong impact on the total variance. It is therefore important

that interviewer effect estimates are prepared so that appropriate survey designs for minimizing the total variance in surveys can be made.

5.4 Required Increase in Budget to Estimate the Interviewer Effect

We have seen that for a fixed budget, increasing the degree of intergroup interpenetration associated with a survey requires a reduction in the sample size given that all other design variables are held constant. This implies an optimal interpenetrated design for the estimation of the interviewer effect under a fixed budget constraint and has implications for the estimation of other components of total survey error such as the sampling variance.

It is therefore useful to be able to calculate by how much the total budget for running the survey would need to be increased in order to accurately estimate the interviewer effect without impacting on the sampling variance. This can be answered through application of Equation (5.14) for a fixed degree of intergroup interpenetration, d_Z, and a variable total budget, C. For example taking the sample size achieved under full confounding in Figure 5.7 of $n = 3334$ and the same design parameters of $n_{int} = 100$, $n_{wk} = 50$, $c_1 = 1$, $c_2 = 4$, $c_3 = 2$, full intragroup interpenetration as specified by the form of $\sum_i b_i$ in (5.8) and finally $a_Z = 88.8$ (see Table 5.4), we can see that designing a fully interpenetrated survey with the same total sample size would require a total budget of at least $C = 25002$. This turns out to be a budget increase of 15002 over the cost required for a fully confounded design to produce a fully interpenetrated design which can be applied to estimate the interviewer effect with the minimum possible variance for our given sample size.

On the other hand if the primary objective of the survey designer is to minimize the sampling variance of the mean estimate derived from the survey while minimizing the variance of the interviewer effect is a secondary priority, the choice between an increase to the total budget and the potential implications on the accuracy of the interviewer effect estimate will depend on the required increase in the total

cost of running the survey. Table 5.6 following expands our example by detailing the increase in budget required to estimate the interviewer effect with a specified variance.

d_Z	vif Variance of $\hat{\sigma}^2_{int}$	Total Cost
0	NA	10000
0.1	10	10962
0.2	5	11922
0.3	3.3	13032
0.4	2.5	14292
0.5	2	15702
0.6	1.7	17262
0.7	1.4	18972
0.8	1.2	20832
0.9	1.1	22842
1	1	$\geq$ 25002

Table 5.6: Variance Inflation Factors Against Total Cost for Fixed Sample Size of 3334

What we can see in Table 5.6 is that the interviewer effect can be estimated with only a small increase in the total budget, however for smaller budget increases the variance inflation factor may still be large. We may be able to make a satisfactory estimate of the interviewer effect for only a modest increase in the total budget. However this will ultimately depend on the priorities of the survey designer.

5.5 Discussion

Our new definition of partial interpenetration combined with modern estimation techniques allows us to estimate the interviewer effect even in surveys with low degrees of interpenetration. In practice almost all surveys contain some degree of interpenetration even if they have not been designed for the purpose of estimating the interviewer effect and this opens up the possibility for widespread application of these techniques. This compares with all current studies for estimating the interviewer effect which have relied on fully interpenetrated survey designs.

The cost of estimating the interviewer effect under partial rather than full interpenetration is however one of accuracy. We have shown that a decrease in the degree of interpenetration leads to an increased variance associated with the interviewer effect estimate. Standard asymptotic variance estimates of the interviewer effect are biased for non-linear response variables and we have therefore introduced a simple empirical framework for producing unbiased estimates of the vifs associated with interviewer effect estimates made under competing survey designs. Based on this we can compare any two competing survey designs with respect to the estimation of the interviewer effect and through this determine the most appropriate survey design. However application of these techniques for the purpose of producing optimal interpenetrating designs for the estimation of the interviewer effect would require prior knowledge of the magnitude of all random effects parameters and in practice this information is generally not available to the survey designer.

We have therefore established a general approximate relationship between the variance of interviewer effect estimate, the sample size and the degree of intergroup interpenetration which can be applied in practice for the optimal design of partially interpenetrating surveys. These general relationships can be applied to produce optimal interpenetrating designs without prior knowledge of the magnitude of the interviewer effect. Although these general relationships are an approximation, they can be used to optimally design partially interpenetrated surveys for the estimation of the interviewer effect based on a budget constraint. This is an important advance in the practical estimation of the interviewer effect as full interpenetration is often too costly in practice but a minor increase in the degree of interpenetration for an existing survey may be cost effective.

An optimal design can be found because increasing the degree of interpenetration leads to increased costs and hence to conduct a survey on a fixed total budget, C, we would have to reduce the sample size to accommodate an increased degree of interpenetration. These increased costs will generally be associated with more interviewers travelling to a greater number of spatial areas as we alter the degree of intergroup interpenetration, thereby increasing the total travel costs for the survey.

The potential impact of estimation of the interviewer effect on other components of the total survey error, such as the sampling variance, under a fixed budget were also explored. The total budget increase required to produce an appropriate estimate of the interviewer effect without affecting the sampling variance was then presented.

In summary we have demonstrated how to produce a valid estimate of the interviewer effect under partial interpenetration and the variance inflation factors that can be associated with these designs. Using these techniques it will generally be possible to produce an appropriate estimate of the interviewer effect with only a minor change to current survey designs and at a low cost.

Further extensions to this work would be to consider more complex travel cost functions and to explore the effect of changing the degree of intragroup interpenetration in more detail. When preparing partially interpenetrated surveys there will be a number of practical considerations which will also influence workload formulation decisions. In this chapter we have demonstrated the potential gain from utilizing partial interpenetration, however the actual gain will depend on the structure of the survey to which it is applied. Fully exploring the implications of a design for a non-linear response is still somewhat computationally intensive, however the approximate relationships presented in this chapter allow rapid calculation of optimal design parameters. Multiple objective optimal designs can be considered in more detail and the relative benefits of greater within time period or between time period temporal interpenetration are still to be explored.

Chapter 6

Spatial Modelling for Interviewer Effects

Traditional methods for estimating the interviewer effect rely on some form of interpenetration. We have demonstrated how the longitudinal information in a repeated panel survey can be utilized to increase the effective degree of interpenetration associated with a survey and therefore improve estimates of the interviewer effect. There is further information available to the survey methodologist that may allow the regional and interviewer effects to be disentangled when a survey is not fully interpenetrated. We will now assess whether there are potential gains from incorporating spatial information into the estimation process.

In this chapter we refer to spatial models as a general class of models which incorporate information regarding the geographical distribution of the data. The variance decomposition models applied in previous chapters only estimated spatial effects by utilizing the classification structure of the data and assumed that all responses within a region exhibited a constant correlation. Consequently these models ignored any remaining spatial autocorrelation in the data. We will extend our earlier variance decomposition techniques to include spatial autocorrelation terms to model the spatial structure of the data. The spatial models applied in this chapter will be similar to those adopted in spatial epidemiology (see for example Richardson and Monfort (2000)) in which spatial models are generally applied to cater for the

spatial autocorrelation structure of the data.

It is appealing to separately model the spatial effect within a variance decomposition model for the estimation of the interviewer effect. This is because any expected spatial effect is not generally a product of a specific spatial zoning structure. This is related to the zoning problem in the spatial modelling literature, in which it is recognized that the spatial boundaries relating to any grouping or zoning structure in a spatial dataset are an arbitrary construct. Consequently significant spatial effects may be estimated for a dataset no matter what spatial grouping structure is applied. For example we might still estimate significant, though different, spatial effects no matter whether the spatial boundaries are postcodes, electoral divisions or suburbs. This compares with the interviewer effect in which the correlation between responses due to the interviewer is linked to a single grouping structure - in this case the allocation of respondents to interviewers.

While interviewer effects can be explained by correlation between responses in a single interviewer's allocation, spatial effects are not linked to a single grouping structure. We may be able to better describe the spatial correlation structure in a variance decomposition model as perhaps a function of the distance between respondents or by alternative methods for considering the relative location of respondents.

This chapter will briefly review spatial modelling techniques and present methods that can be applied to estimate the interviewer effect. An application will then be presented based on both repeated panel and single month examples drawn from a unit record file derived from the Australian Bureau of Statistic's (ABS's) Monthly Population Survey (MPS). It is hoped that explicitly modelling spatial correlation within a variance decomposition model will explain some of the spatial correlation that would otherwise be confounded with the interviewer effect, in these partially interpenetrated scenarios, leading to improved estimates of the interviewer effect.

6.1 Spatial Modelling

Spatial correlation is correlation between units that is in some way related to the position or relative position of the units. This is a generalization of the clustering discussed in the Multiple Membership Multiple Classification (MMMC) models of Browne *et al.* (2001) in which correlation between observations was limited to points in the same classification; group or zone. Spatial modelling refers to models that can be applied to incorporate information regarding the position, or relative position, of observations in our dataset. This can be done either by incorporating spatial covariates, that is information about the spatial regions as a covariate, or we can explicitly model the spatial correlation structure of the data. In spatial correlation models there is no requirement for the correlation between observations in a single cluster to be constant and correlation can extend beyond the cluster boundaries. Consequently spatial modelling is more general than the MMMC class of models. This allows us to explicitly describe any spatial correlation and may reduce confounding between any unexplained spatial correlation in the interviewer allocations and the correlation due to the interviewer. However, spatial models can be much more complex and difficult to estimate than the majority of models in the MMMC class.

There are a wide range of spatial modelling tools commonly applied in practice and these are generally dependent on the amount and form of the spatial data available to the analyst. The following section will introduce the dataset that will be examined and describe the spatial information that is available.

6.1.1 Available Spatial Information

The Confidentialised Unit Record File (CURF) introduced in Chapter 3 was prepared as a sample of 50 workloads from the Labour Force Survey component of the Monthly Population Survey (MPS). To maintain the privacy of respondents a number of data items were aggregated, randomized or removed from the original unit record file sample until it was determined that the disclosure risk from the CURF

was minimal. Due to this confidentialisation procedure all spatial identifiers were removed and consequently spatial modelling could not be performed on the CURF.

The original unit record file sample from the MPS, known as the pre-CURF dataset, from which the CURF was derived, was maintained at the ABS. The spatial information available on the pre-CURF is the northing and easting grid positions of the centroid of the Collection Districts. Northings and eastings are projections of alternative measures, such as longitude and latitude, onto a cartesian surface known as the Map Grid of Australia 1994 (MGA94). On-site access to this unconfidentialised file was granted for the purposes of this project. Spatial models for the estimation of the interviewer effect were fitted on the pre-CURF at the ABS with the understanding that only the aggregate results from this procedure could be reported. Consequently diagnostics associated with spatial models, including residual plots, have not been presented in this chapter. Where relevant these diagnostic tools were utilized and subsequent observations have been recorded.

6.1.2 Spatial Modelling Review

Research into spatial statistics has expanded along with access to more powerful computers and the development of Geographic Information Systems (GIS). Much of the work in spatial statistics has followed a data driven spatial analysis approach to produce either global or local measures of spatial correlation for a dataset. On the other hand, spatial econometric modelling is interested in explicitly modelling the spatial autocorrelation of observations within a dataset. Early work into the modelling of spatial correlation recognized that spatial dependence between observations tended to diminish over increased distance. Consequently time series techniques were adapted to the analysis of spatial data in order to cater for this spatial autocorrelation. Spatial autocorrelation can be considered to be even more complex than temporal autocorrelation as observations tend not to be equally spaced and the autocorrelation exists in more than one dimension. Some adaptations to deal with this increased complexity have been considered, such as incorporating spatial weight matrices to reflect the relative position of observations and geographically

weighted regression.

Common autoregressive models currently employed in spatial analysis are the spatial Conditional AutoRegressive (CAR) model and the spatial Simultaneous AutoRegressive model (SAR). These models are specified slightly differently and the CAR is the most appropriate formulation for considering first order dependency while the SAR can be used to also consider second order dependencies. Bao (2001) highlights that the spatial SAR model is more prevalent in spatial studies, which generally consider both first and second order dependencies. These models allow specification of the spatial autocorrelation of observations but are not overly complex. Reviews of spatial analysis can be found in Bao (2001) and Bivand (1998), examples of the application of spatial CAR and SAR models can be found in Breslow and Clayton (1993) and Case (1991) respectively.

The remainder of this chapter will consider simple variance decomposition models for the estimation of the interviewer effect. Given the complex hierarchical spatial structure in the pre-CURF dataset (see Figure 3.7), spatial CAR models will be applied to the pre-CURF dataset to model the first order spatial autocorrelation between Collection Districts (CDs).

Spatial CAR and Spatial MM Models

Spatial CAR models can be considered a subset of spatial multiple membership (MM) models. Spatial MM models assume that regions neighbouring the region to which an observation belongs can still be considered to influence that observation. Given a weighting structure outlining the relative contribution of the multiple spatial areas, spatial MM models can be applied to relax the assumption that each observation can be considered part of a single spatial area. In order to limit the complexity of this class of models, observations are usually only considered to have multiple membership within neighbouring spatial regions. However this is not a strict requirement and a neighbourhood can be defined in a number of ways. For the purposes of this chapter we will consider that two CDs are neighbours if they are within a specified distance, d_s, of one another. Consequently neighbouring CDs

do not have to literally share a boundary, but they do have to be within close spatial proximity to one another.

In practice spatial MM models fit a random effect for each neighbouring spatial region. These random effects are then linked via a distributional assumption. An example of the application of a spatial MM model, and a comparison with the spatial CAR model on the Scottish lip cancer dataset can be found in Browne (2002). Variance components relating to spatial MM terms can be represented as a standard random effect under the MMMC representation of GLMMs presented in Chapter 2. However to clarify the form of the spatial MM term we can adopt the notation of Browne *et al.* (2001) to demonstrate the form of the term in the variance decomposition model to be fitted to the pre-CURF. For example

$$\Pr(y_{ijklqmn} = 1|\pi_{jklqmn}) = \frac{\exp(\pi_{jklqmn})}{1 + \exp(\pi_{jklqmn})} \tag{6.1}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the CURF.

$$\pi_{jklmn} = \mu + \omega_n + \nu_{mn} + MM_{qmn} + \varphi_{lqmn} + \phi_{klqmn} + \theta_{jlkqmn}$$

and

- μ is a fixed effect
- The multiple membership term, MM_{qmn} can be written under the notation of Browne *et al.* (2001) as $\sum_{q \in neigh(i)} w^{(5)}_{i,q} u^{(5)}_{1,q}$ where w is a weight describing the relative contribution of the neighouring regions. This clarifies that MM_{qmn} is a variance component at the fifth level in the classification structure relating to the multiple membership of observations, i, in neighbouring spatial regions. This term is formed as the weighted sum of random effects for the neighbouring regions and we assume that the random effect from which this sum is formed is normally distributed, i.e. $u^{(5)}_{1,q} \sim N(0, \sigma^2_u)$
- The remaining random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_{int})$ and $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$

- i, j, k, l, q, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, the CD neighbourhood, interviewer and workload levels respectively
- The classification structure of the data can be seen in Figure 3.7

Spatial CAR models set up an autoregressive model to describe the spatial correlation in neighbouring regions. Consequently under the spatial CAR model only one parameter needs to be estimated per spatial region in order to describe all of the neighbouring spatial regions. The spatial CAR and the spatial MM models are broadly consistent. However spatial CAR models are less complex and therefore tend to be more stable. A spatial CAR model for employment status on the pre-CURF dataset can be written as

$$\Pr(y_{ijklqmn} = 1|\pi_{jklqmn}) = \frac{\exp(\pi_{jklqmn})}{1 + \exp(\pi_{jklqmn})} \tag{6.2}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the CURF.

$$\pi_{jklmn} = \mu + \omega_n + \nu_{mn} + CAR_{qmn} + \varphi_{lqmn} + \phi_{klqmn} + \theta_{jlkqmn}$$

and

- μ is a fixed effect
- The spatial CAR term, CAR_{qmn} can be written under the notation of Browne *et al.* (2001) as $CAR_{qmn} \sim N\left(\sum_{j \in neigh(i)} \frac{\theta_j}{r_i}, \frac{\sigma_v^2}{r_i}\right)$ which clarifies that CAR_{qmn} is a variance component relating to the conditional autoregressive relationship of observations, i, in neighbouring spatial regions
- The remaining random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma_\omega^2)$, $\nu_{mn} \sim N(0, \sigma_\nu^2)$, $\varphi_{lmn} \sim N(0, \sigma_\varphi^2)$, $\phi_{klmn} \sim N(0, \sigma_{int}^2)$ and $\theta_{jklmn} \sim N(0, \sigma_\theta^2)$
- i, j, k, l, q, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, the CD neighbourhood, interviewer and workload levels respectively

- The classification structure of the data can be seen in Figure 3.7

The spatial CAR model is not often applied to binary data items in practice. This is because the sum of a number of binomial random variables is not itself binomial and hence the binomial formulation, as in Model (6.2) does not aggregate conveniently to higher levels. Wakefield *et al.* (2000, p 106) recognize that this is not likely to cause a problem,

'*In principle, this lack of data reduction is not a problem. However it may create computation difficulties due to memory requirements if n* (the sample size) *and/or J* (the number of strata) *are large and, in practice, numerical estimation problems may also arise if there are large numbers of (i,j) cells containing zero cases since in this case the likelihood is likely to be very flat and hence contain little information.*' We have seen in Chapter 3 that there are no zero cells in the pre-CURF. Consequently it should be valid to apply both the spatial CAR and the spatial MM models in this case.

The following section will examine the pre-CURF to determine which CDs are neighbours and to develop an understanding as to the possible extent of the spatial correlation for employment status. Later sections will apply both the spatial CAR Model (6.2) and the spatial MM Model (6.1) to explicitly model the spatial effect.

6.2 Pre-CURF Dataset

Before complex spatial models were applied to the pre-CURF the dataset was examined to determine whether there was evidence of spatial correlation in the data. As in Chapter 3 a binary employment status variable was the data item of primary interest for this analysis.

We have already seen in Chapter 3 that there was a significant CD effect for binary employment status. In the absence of other covariates which may explain this correlation, this suggests that there is some spatial correlation at the CD level in the CURF dataset. On the pre-CURF northings and eastings of the CD centroids is available. Based on MGA94 these co-ordinates are arranged on a cartesian grid

and hence the distance, in meters, between any two CDs can be calculated using the following formula

$$d_{CD1,CD2} = \sqrt{(Northing_{CD1} - Northing_{CD2})^2 + (Easting_{CD1} - Easting_{CD2})^2} \tag{6.3}$$

Based on expression (6.3) a matrix was then formed specifying the distance between each pair of CDs.

6.2.1 Determining Neighbouring CDs

Given our matrix specifying the distance between any pair of CDs, neighbouring CDs were defined as being within a specified distance, d_s, of one another. The choice of an appropriate distance, d_s, within which CDs could be considered as being neighbours was then a trade-off between the complexity of increased numbers of neighbouring CDs and the reduced applicability of small numbers of neighbours to explain the spatial correlation. In particular for a spatial MM model, specifying high numbers of neighbouring CDs would greatly increase the complexity and reduce the stability of the model as the number of random effects to be estimated increases along with the total number of neighbours.

In the pre-CURF there are 387 CDs and 74,691 CD to CD combinations between pairs of CDs. A distance cut-off of $d_s = 8,000$ meters was chosen such that 2,076 CD to CD pairs were found to be within this distance of one another. Consequently approximately 0.7% of all possible CD to CD combinations were considered to be within a neighbouring distance of one another. The maximum number of neigbhouring CDs to any single CD was found to be 15. Although the actual distance between CDs could be used to formulate weights to specify the potential contribution of each neighbour to the estimated random effect, for simplicity it was assumed that each neighbouring CD would be attributed an equal weight.

Based on this information columns were attached for each observation in the pre-CURF dataset identifying neighbouring CDs and the weights that could be

associated with each these neighbouring CDs.

This process was then repeated in order to determine neighbouring workloads. The position of the workloads was estimated as the mean northing and easting of all observations within that workload. Although the position of the observations was only available at the CD level, this approach has the advantage of weighting the influence of the CDs on the estimated workload northing and easting according to the number of observations within each CD.

In the pre-CURF there are 50 workloads and 1,225 workload to workload combinations between pairs of workloads. A distance cut-off of $d_s = 100,000$ meters was chosen and 92 workload to workload pairs were found to be within this distance of one another. Consequently approximately 7.5% of all possible workload to workload combinations were considered to be within a neighbouring distance of one another. The maximum number of neigbhouring workloads to any single workload was found to be 8. Equal weights were again assumed to consider the relative influence of neighbouring workloads.

6.2.2 Exploration of the Data

The pre-CURF dataset was briefly examined to determine whether there was any evidence of spatial correlation in employment status before more complex spatial models were fitted to the data. We have already seen in Chapter 3 that there is a significant CD level effect in employment status and this suggests that there is some spatial correlation in the data. However, given that the workload and the interviewer effects are small in comparison, this does not tell us whether the spatial correlation extends much beyond the CD boundaries.

As we have defined spatial correlation as correlation between observations in our dataset that is somehow related to the position or relative position of our data we can get an initial idea as to the approximate extent of the spatial correlation in our data by simply fitting some basic functional descriptions of the position of the observations in our dataset. For example Tagashira and Okabe (2002) describe models which consider explaining some of the spatial correlation by fitting the dis-

tance from a pre-determined point as an explanatory variable. We have adopted this approach in our initial examination and have chosen the CD with the highest number of neighbours (henceforth referred to as CD^*) as the baseline CD against which a distance d^* has been calculated. As a single reference point CD^* provides us with the highest number of neighbours and therefore observations in this CD will be correlated with the highest number of other observations. Thus for any CD,

$$d^* = \sqrt{(Northing_{CD} - Northing_{CD^*})^2 + (Easting_{CD} - Easting_{CD^*})^2} \tag{6.4}$$

When we examine the CDs we can see that there are strong differences in the proportion of employed by CD. This information is presented in Figure 6.1 following. This reinforces the significant CD level effect that was estimated for Model (3.7) in Chapter 3 and suggests that there are significant spatial effects at the CD level for employment status.

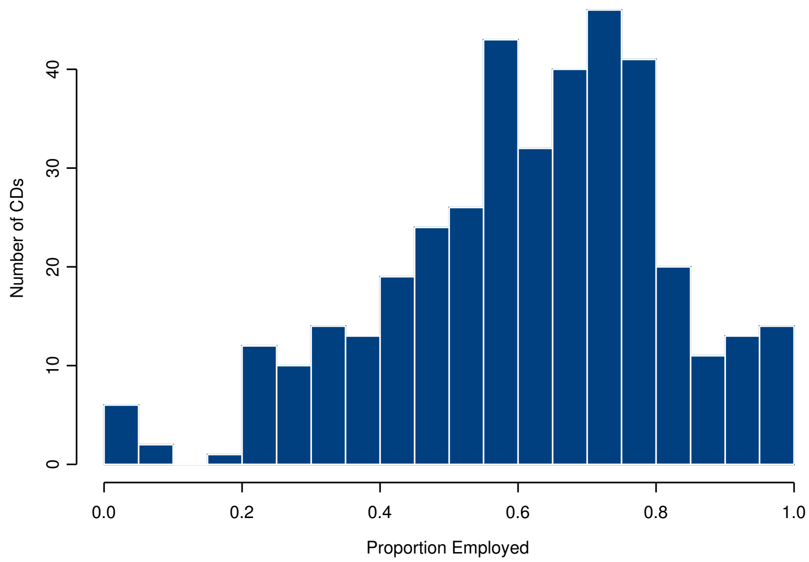

Figure 6.1: Histogram of Proportion of Employed in Each CD

Figure 6.1 does not clarify the extent of the spatial correlation and we have already seen in Model (3.7) that the majority of the spatial effects under the MMMC variance

decomposition model appear below, rather than above, the CD level. Consequently we might not expect a spatial autocorrelation term at the CD level to explain much of the correlation in the data.

We can further examine the possible impact of spatial correlation on employment status by assessing the relationship of the distance between observations on employment status. This approach suggests that respondents that live closer to one another will exhibit greater spatial correlation. We begin by calculating the correlation between employment status and a number of spatial variables.

	D_{nth}	D_{est}	d^*	Emp Stat	Nth	Est
D_{nth}	1	*	*	*	*	*
D_{est}	0.38	1	*	*	*	*
d^*	0.99	0.44	1	*	*	*
Emp Stat	0.04	0.03	0.04	1	*	*
Nth	0.93	0.38	0.92	0.04	1	*
Est	0.03	0.31	0.06	0.03	0.03	1

Table 6.1: Correlation Between Employment Status and Spatial Variables

In Table 6.1, Nth stands for northing, Est for easting and Emp Stat for employment status. In this case D_{nth} and D_{est} are the distances from CD^* in the longitude and latitude directions respectively so that $D_{nth} = \sqrt{(Northing_{CD1} - Northing_{CD^*})^2}$ and $D_{est} = \sqrt{(Easting_{CD1} - Easting_{CD^*})^2}$. We can see in Table 6.1 that most of the separation between CDs in the dataset occurs along the northing axis.

We will examine whether there appears to be a relationship between the distance of observations from CD^* and the employment status of individuals in order to assess whether this appears to be a good model to describe the spatial correlation. An inverse exponential function of the distance, d^*, was considered as observations that are further apart generally exhibit weaker spatial correlation.

The following regression model was fitted to employment status on the pre-CURF

$$\Pr(y_{ij} = 1|\pi_j) = \frac{\exp(\pi_j)}{1 + \exp(\pi_j)} \tag{6.5}$$

where the probability of being employed or unemployed π_j is a function of the

distance of the observation's CD from the baseline CD, such that

$$\pi_j = \mu + \frac{\beta}{\exp(d_j^*)}$$

where for each response i within CD j, d_j^* is the distance of the respondent's CD from the baseline and y_{ij} is the observed employment status. The estimated parameters for Model (6.5) are $\mu = -0.511(0.016)$ and $\beta = -0.627(0.102)$. The significant spatial effect estimate in Model (6.5) suggests that there is some CD level spatial correlation in the data that can be modelled as a function of the distance between CDs. This is as expected as under Model (3.7) we had estimated a significant CD level effect. However, the Deviance Information Criterion (DIC) statistics of Spiegelhalter *et al.* (2002) suggest that incorporating a term for the distance between observations has barely increased the explanatory power of the model. This can be seen if we compare Model (6.5) with a constant only model

$$\Pr(y_i = 1|\pi) = \frac{\exp(\pi)}{1 + \exp(\pi)} \tag{6.6}$$

where the probability of being employed or unemployed π is a constant for all observations, i, i.e. $\pi = \mu$ and for each response i, y_i is the observed employment status.

The DIC for Model (6.6) is 24000, while the DIC for Model (6.5) is also 24000 to two significant figures. This indicates that there has been very little change to the explanatory power of the model from including an explanatory variable for the distance between observations, even though this term is significant. Part of the reason for this is that in Model (6.5) we are only attempting to describe the correlation between observations with respect to the distance from a single CD. We have already seen that only 15 CDs can be considered to neighbour this CD, so consequently our explanatory variable does not have a marked overall impact on the explanatory power of the model.

6.2.3 Summary

We have explored the data briefly to examine whether there is evidence of spatial correlation in the employment status variable on the pre-CURF. We had seen in Chapter 3 that there was a significant CD level spatial effect on the data and in this section we presented a basic analysis to determine whether the data appeared to be spatially correlated.

In the absence of any other variance components we determined that the distance between CDs is a significant explanatory variable for employment status. It is likely that the impact of this explanatory variable would be reduced in the presence of the remaining variance components highlighted under Figure 3.7.

In summary we have seen that there is evidence of spatial correlation at the CD level for employment status in the pre-CURF and that this correlation can be modelled using spatial techniques. The spatial correlation appears to exist mainly below the CD level and since the available spatial information is the northing and easting of CD centroids, modelling of the spatial correlation as a function of distance will not be pursued in more detail. Instead spatial CAR and spatial MM models will be applied to consider the relationship between observations in neighbouring CDs and workloads.

6.3 Results

In this section the spatial CAR Model (6.2) and the spatial MM Model (6.1) are fitted to the employment status data item on the pre-CURF to determine whether explicitly modelling the spatial correlation in the data may aid in the estimation of the interviewer effect.

6.3.1 Application of Spatial CAR and Spatial MM Models

The spatial CAR Model (6.2) and the spatial MM Model (6.1) were fitted using the MCMC estimation algorithm available in MLwiN 2.0. The steps that were followed for determining appropriate priors, starting values and convergence are discussed in

Section 3.4.2. Variance component estimates for these models can be seen in Table 6.2 below.

	Spatial Car model (6.2)		Spatial MM Model (6.1)	
Variance Component	Estimate	Standard Error	Estimate	Standard Error
Workload	2.49	2.05	3.53	2.44
Interviewer	0.11	0.09*	0.09	0.10*
Neighbouring CDs	3.17×10^{-5}	6.58×10^{-5}	0.05	0.05*
CD	15.7	4.35	16.0	3.95
Household	82.1	9.36	85.3	10.2
Person	105.2	10.3	109.8	10.7
Fixed Effect	3.73	0.42	3.86	0.45

Table 6.2: Parameter Estimates for Binary Employment Status, Spatial CAR Model (6.2) and Spatial MM Model (6.1) for Repeated Panel Data

The * in Table 6.2 indicates estimates for which the correlation in the MCMC chain is high and for which the posterior trace does not appear to be normally distributed.

The standard errors presented in Table 6.2 are Bayesian equivalences available under MCMC estimation. They are derived from the estimated posterior and hence will be equivalent in interpretation to a frequentist standard error provided the posterior is normally distributed. Due to the complexity of the models in Table 6.2 there is a high degree of correlation between successive iterations of the MCMC chain and consequently the estimation procedure was halted before all parameters appeared to have appropriately converged to the posterior. In this case results for both the spatial CAR Model (6.2) and the spatial MM Model (6.1) are presented based on a MCMC chain of 100,000 iterations. The Brooks-Draper statistic gives an indication that neither the spatial CAR Model (6.2) nor the spatial MM Model (6.1) has converged appropriately to the posterior and that the MCMC chain should be run for further iterations for convergence to be achieved. For example the Brooks-Draper statistic indicates that we would need to run the MCMC chain for more than 7,500,000 iterations for the spatial MM term to be estimated to two significant figures in the spatial MM Model (6.1). In comparison the spatial CAR term in Model (6.2) appears to have already appropriately converged to the posterior, with

a Brooks-Draper of approximately 7,000. However, the interviewer effect term for the spatial CAR Model (6.2) is still to achieve convergence with a Brooks-Draper of 500,000. Along with the high autocorrelation in the MCMC chain affecting some of the estimates, examination of the diagnostic traces indicated that some of the estimated posteriors were not normally distributed. In each case this occurred for parameters with highly autocorrelated MCMC chains. Indicative Bayesian standard errors have still been presented for these parameters.

We can see in Table 6.2 that the parameter estimates for all of the variance components, ignoring the CAR and MM term, are very similar to estimates presented in Table 3.12 for the standard variance decomposition model (3.2). The estimated spatial MM and CAR terms are also of low magnitude in comparison to the remaining variance components and are non-significant. Furthermore if we consider the DIC we can see that inclusion of the spatial MM and CAR terms does not appear to explain much of the variation in binary employment status on the CURF.

In summary the spatial CAR Model (6.2) appears to be more stable than the spatial MM Model (6.1) when estimating interviewer effects on the employment status data item in the pre-CURF. This is likely due to the reduced complexity of the spatial CAR formulation in comparison to the spatial MM model, as the remaining variance decomposition structure in this case is already quite complex. This is indicated by the relative levels of autocorrelation in the two models. However even for the spatial CAR Model (6.2) convergence was not achieved in the limited time available to fit these models. Based on the results presented in Table 6.2 we can see that the estimated spatial CAR term does not appear to explain much of the variation in the data. Consequently explicit spatial modelling of the spatial correlation does not appear to aid in the estimation of the interviewer effect in this case and may actually inhibit estimation due to the increased complexity of the model.

6.3.2 Spatial Modelling with Low Interpenetration

We have seen in Table 6.2 that explicit spatial modelling of the relationship between neighbouring CDs does not appear to explain much of the spatial correlation for employment status in the pre-CURF. Part of the reason for this is that it appears most of the spatial correlation in the data exists below the CD level and hence we might not expect much gain from allowing the correlation to cross CD boundaries. Our aim however was to see if we could use spatial modelling techniques to explain some of the spatial correlation that might otherwise contribute to the confounding between the interviewer and spatial effects. With regards to this aim, applying the spatial CAR and spatial MM models to the full pre-CURF data would also not provide much gain as the full pre-CURF has a high effective degree of interpenetration when treated as a longitudinal dataset. Furthermore the confounding between interviewer and spatial effects would be expected at the workload rather than the CD level at which we modelled the spatial effect in Model (6.2). Consequently we would not expect to see much of an improvement in estimation of the interviewer effect due to a reduction in the confounding spatial effects for the full pre-CURF dataset. If we consider a single month of the pre-CURF dataset and explicitly model the spatial autocorrelation between neighbouring workloads we might expect to see a greater impact on the estimation of the interviewer effect. This will also simplify the spatial CAR and MM models as the number of neighbouring regions will be reduced. Consequently for the remainder of this chapter we will explore the applicability of spatial MM and CAR models for describing spatial autocorrelation at the workload level in a scenario with low interpenetration, in this case the single month scenario.

A single month, November 2001, was chosen from the pre-CURF dataset. Each respondent was only interviewed once in November and each household was only interviewed by a single interviewer, so consequently the classification structure of the data is simpler in a single month with one fewer level than the full pre-CURF. The classification structure of the data can be seen in Figure 3.4. A new formulation of the spatial MM model was set up to reflect this new classification structure

$$\Pr(y_{jklmnq} = 1 | \pi_{klmnq}) = \frac{\exp(\pi_{klmnq})}{1 + \exp(\pi_{klmnq})} \tag{6.7}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the CURF.

$$\pi_{klmnq} = \mu + MM_q + \omega_{nq} + \nu_{mnq} + \varphi_{lmnq} + \phi_{klmnq}$$

and

- μ is a fixed effect
- The multiple membership term, MM_q can be written under the notation of Browne *et al.* (2001) as $\sum_{q \in neigh(j)} w^{(6)}_{j,q} u^{(6)}_{1,q}$ which clarifies that MM_q is a variance component at level 6 relating to the multiple membership of respondents, j, in neighbouring spatial regions. This term is formed as the weighted sum of random effects for the neighbouring regions and we assume that the random effect from which this sum is formed is normally distributed, i.e. $u^{(6)}_{1,q} \sim N(0, \sigma^2_u)$
- The remaining random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$ and $\phi_{klmn} \sim N(0, \sigma^2_{int})$
- j, k, l, q, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, the workload neighbourhood, interviewer and workload levels respectively
- The classification structure of the data can be seen in Figure 3.4

while a spatial CAR model for a single month of the pre-CURF dataset is

$$\Pr(y_{jklmnq} = 1 | \pi_{klmnq}) = \frac{\exp(\pi_{klmnq})}{1 + \exp(\pi_{klmnq})} \tag{6.8}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the CURF

$$\pi_{jklmn} = \mu + CAR_q + \omega_{nq} + \nu_{mnq} + \varphi_{lmnq} + \phi_{klmnq}$$

and

- μ is a fixed effect
- The spatial CAR term, CAR_q can be written under the notation of Browne *et al.* (2001) as $CAR_q \sim N\left(\sum_{q \in neigh(j)} \frac{\theta_q}{r_j}, \frac{\sigma_v^2}{r_j}\right)$ which clarifies that CAR_q is a variance component relating to the conditional autoregressive relationship of respondents, j, in neighbouring spatial regions
- The remaining random effects are all independent and normally distributed, i.e. $\omega_n \sim N(0, \sigma_\omega^2)$, $\nu_{mn} \sim N(0, \sigma_\nu^2)$, $\varphi_{lmn} \sim N(0, \sigma_\varphi^2)$ and $\phi_{klmn} \sim N(0, \sigma_{int}^2)$
- j, k, l, q, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, the workload neighbourhood, interviewer and workload levels respectively
- The classification structure of the data can be seen in Figure 3.4

The spatial CAR Model (6.8) and the spatial MM Model (6.7) were again fitted using the MCMC estimation algorithm in MLwiN 2.0. The steps that were followed for determining appropriate priors, starting values and convergence are discussed in Section 3.4.2. Variance component estimates for these models can be seen in Table 6.3 below.

	Spatial Car model (6.8)		Spatial MM Model (6.7)	
Variance Component	Estimate	Standard Error	Estimate	Standard Error
Neighbouring Workloads	1.9 $\times 10^{-3}$	2.7 $\times 10^{-3}$	0.03	0.06*
Workload	0.05	0.06	0.05	0.06
Interviewer	0.07	0.06*	0.07	0.07*
CD	0.41	0.12	0.41	0.12
Household	2.03	0.29	2.02	0.29
Fixed Effect	-0.65	0.08	-0.65	0.08

Table 6.3: Parameter Estimates for Spatial CAR Model (6.8) and Spatial MM Model (6.7) in November 2001

The * in Table 6.3 indicates estimates for which the autocorrelation is high and for which the posterior trace does not appear to be normally distributed.

The magnitude of the variance component estimates at levels which are common for both Table 6.2 and Table 6.3 are different. This is due to observations within

respondents being the lowest level in the multiple month case while the respondents are the lowest level in the classification structure in the single month scenario. We have already seen that most of the variation exists between respondents rather than within respondents over time, so consequently the estimated variance components are very different in the single and multiple month case. Although the number of spatial areas, workloads and CDs, are similar for both the single month and the multiple month case, the number of interviewers in any single month is less than half that of the longitudinal case. Similarly there are fewer respondents available in any single month - see Chapter 3 for more details. Thus, although we might expect the interviewer level variance component estimates to be similar for both the single and multiple month scenarios, the body of interviewers upon which the interviewer effect estimates are based will be different. If we then also consider any potential effect of the degree of interpenetration there may be substantial variation between estimates of the interviewer effect produced from either the single or the multiple month case.

We have seen that the classification structure of the data is simpler in the single month case and consequently we are able fit simpler spatial models to describe the data. However there is also less data available in the single month case. Both the spatial CAR Model (6.8) and the spatial MM Model (6.7) were run for 250,000 iterations of the MCMC chain, but diagnostics indicated that the chains had yet to converge appropriately to the posterior. The Brooks-Draper was highest for the interviewer effect estimate on the spatial CAR Model (6.8), indicating that the MCMC chain should be run for 1,500,000 iterations to produce an estimate to two significant figures. The more complex spatial MM Model (6.7) was again less stable with higher autocorrelation in the MCMC chain, indicating that the spatial CAR model was again preferred in this case.

In summary there was again little benefit from explicitly modelling the spatial autocorrelation structure between workloads. Even though there is a much lower degree of interpenetration in the single month case, the reason for this again appears to be that most of the spatial correlation for the employment status data item in the

pre-CURF exists below the CD level. Consequently even though the data structure is almost fully confounded in the single month case, explicit spatial modelling does not appear to help us explain much of the workload level effect and subsequently reduce the confounding.

6.4 Discussion

In this chapter we applied spatial models to explicitly describe the spatial correlation between observations in the pre-CURF. It was hoped that in this way some of the workload level spatial effect that might otherwise confound estimation of the interviewer effect could be explained.

In practice it appears that most of the spatial correlation in the pre-CURF dataset exists below the CD level. As the potential confounding was expected to primarily be between interviewer allocations and the workloads, explicit modelling of the spatial autocorrelation structure between either CDs or workloads did not lead to an improvement in estimates of the interviewer effect. Furthermore the increased complexity of the spatial models over simpler variance decomposition models actually inhibited our ability to estimate the interviewer effect in practice.

Although spatial modelling of the pre-CURF dataset has not lead to improved estimates of the interviewer effect this is partly due to the survey design and the degree of spatial correlation for our data item of interest. The increased complexity of these spatial models means that they should be applied with care. However in certain circumstances they may still be beneficial for the estimation of the interviewer effect. For example if the classification structure of the data was less complex, the inclusion of a spatial autocorrelation term may be viable. Furthermore when spatial correlation is evident at the confounded level we might expect these techniques to produce greater benefits.

Chapter 7

Conclusion

The role of the interviewer in households surveys is multi-faceted. The interviewer may be required to elicit and prompt response, collect and enter data and also provide initial contact with respondents. Interviewers therefore play a central role in the collection of high quality data. Besides these positive influences, the presence of the interviewer may also have an unintended impact on survey data. In particular it has been noted that responses collected by the same interviewer tend to be more similar than would otherwise be expected if the responses were collected by different interviewers. The presence of the interviewer may therefore lead to an unintended correlation in survey data and an increased uncertainty in responses the interviewer collects. We must therefore estimate this interviewer effect on total survey error so that results derived from surveys can be applied appropriately.

Previous studies examining the interviewer effect have generally relied on fully interpenetrated designs, in which a minimum of two interviewers are randomly allocated to each spatial area. Interpenetration provides repeated measurement of spatial areas and allows the spatial and interviewer effects to be disentangled. However, conducting an interpenetrated survey is an expensive process and does not occur often in practice.

This thesis has taken the first steps towards establishing a comprehensive framework for the estimation of the interviewer effect. The potential gain from incorporating any longitudinal and spatial information available to the survey designer has

been considered and methods for estimating the interviewer effect in partially interpenetrated surveys have been introduced. Optimal partially interpenetrated designs have been prepared that allow more cost effective estimation of the interviewer effect in practice. These advances greatly increase our ability to estimate the interviewer effect in a much wider range of practical scenarios and will lead to more frequent estimation of the interviewer effect.

7.1 Conclusions

There have been a number of innovations introduced in this thesis. Most notably we have defined interpenetration comprehensively and introduced the concept of partial interpenetration for the first time. We have considered how interviewer effects can be estimated in practice for large repeated panel household surveys and demonstrated how spatial and longitudinal information can be incorporated into this process. Some specific outcomes developed in this thesis are

- Unless you have a reason to suspect that some interviewers are exceptional, such as receiving special training, there is generally no need to conduct a fully interpenetrated survey. Partially interpenetrated survey designs are generally more cost effective than fully interpenetrated designs for the estimation of the interviewer effect.

- Optimal partially interpenetrated designs can be prepared by recognizing that for a fixed total budget there is a trade-off between increased sample size and increased degrees of interpenetration.

- Provided interviewers are rotated over time, longitudinal information, such as that available in repeated panel surveys can be used to greatly increase the effective degree of interpenetration associated with a survey and therefore improve the interviewer effect estimate.

- Approximate methods for estimating the interviewer effect, such as MQL and PQL, may be strongly biased for discrete data items in household surveys. This

occurs because there are generally only a small number of responding persons within each household or a small number of observations per respondent and the bias is related to the number of first level units within the second level units in the classification structure.

- Exact methods for estimating the interviewer effect should be used for discrete data items when the expected bias from approximate estimation methods is large. The expected bias can be demonstrated using simulation techniques presented in this thesis. Currently MCMC techniques are the most flexible of the exact methodologies for the estimation of the interviewer effect.
- The impact of the interviewer effect will generally be minimized when interviewers are allocated equal numbers of respondents.

This thesis has demonstrated how the interviewer effect can be estimated under a wide variety of survey designs that were previously considered confounded. In some cases these techniques could be applied directly to existing surveys as part of the survey process with little change in the survey design. Regular estimation of the interviewer effect will be a positive advance for the ongoing quality monitoring of large scale surveys and for questionnaire design. In particular, questions which are associated with high interviewer effects can now be rapidly identified and corrected. Furthermore estimation of the interviewer effect will lead to more appropriate interviewer training and debriefing. This is an advance over traditional interpenetrating designs for the estimation of the interviewer effect that are generally too expensive to apply in practice.

7.2 Further Research

Estimation of the interviewer effect will provide users with a much clearer representation of the level of certainty they can associate with results derived from surveys and provide a new framework of tools as part of the ongoing monitoring of large-scale surveys. The interviewer effect is only one component in the decomposition

of total survey error and further research is still needed into workable total survey error models. In this thesis we have demonstrated how to estimate the interviewer effect on a Confidentialised Unit Record File (CURF) and produced a simple total survey error estimate by combining this with a sampling error estimate. Further research into a total survey error framework would help to automate and generalize this process.

The results in this thesis can be naturally extended to consider how the characteristics of the interviewer relate to the interviewer effect. In a similar way the effect of interviewers on non-response, in telephone surveys and on subjective data items can also be examined. Data to explore these issues was not available in the CURF and hence a future study to explore these issues should be considered. Estimation of interviewer effects in other datasets will also require consideration of alternate sampling schemes and survey designs. Probability weighting schemes (see Pfeffermann *et al.*, 1998) can also be incorporated, however preliminary analysis suggests that this technique does not have a strong impact on results.

Further studies, examining the potential gain from incorporating spatial information should also be pursued. We have seen that there was no gain from incorporating the spatial information available in the pre-CURF. This was partly explained by the majority of the spatial correlation existing below the workload level in this dataset. It is likely that explicit modelling of the spatial correlation in the data may lead to further improvements in the estimation of the interviewer effect when the interviewer allocations are very spatially concentrated, however this issue remains to be explored.

7.3 Summary

We have demonstrated how the interviewer effect can be estimated in a variety of new scenarios, including survey designs that were previously considered confounded. The introduction of partial interpenetration means that the interviewer effect and its contribution to total survey error can now be estimated more efficiently and cost-

effectively, potentially as a regular part of the survey process. Regular estimation of the interviewer effect will have positive implications for the ongoing monitoring of large scale surveys and will lead to more appropriate interviewer training, questionnaire design and more cost effective surveys.

Appendix A

Dataset Simulation Code

The following S-Plus code was designed to produce simulated response variables with known effects on the CURF dataset.

A.1 Normal Response Simulation Code: All Months

A response with known random effects was simulated as a new column on the CURF dataset. The CURF level identification labels were used to additively incorporate known random effects to create this simulated variable. The following additive model was used

$$y_{ijklmn} = \mu + \omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn} + \varepsilon_{ijklmn}$$

where

- μ is a fixed effect
- The random effects are all normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_\phi)$, $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$ and $\varepsilon_{ijklmn} \sim N(0, \sigma^2_\varepsilon)$
- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively

The following S-Plus code was then used to simulate this normal response variable onto the CURF structure with known random effect sizes at each of the levels.

```
#set up effect sizes
 gmean_10
 elev1_sqrt(4)
 elev2_sqrt(2)
 elevhh_sqrt(1)
 elevcd_sqrt(0.75)
 elevint_sqrt(0.2)
 elevwkld_0.5
 #data is new3tcurf
 #create column for output - 35
new4tcurf_cbind(new3tcurf[,1:34],rep(gmean,length(new3tcurf[,1])))
 #Person index 28
 #hhindex 27
 #cdindex 31
 #Int index 29
 #Wkld index 30
 #simulate random effect vectors
 tmp2_rnorm(max(new4tcurf[,28]),mean=0,sd=elev2)
 tmp3_rnorm(max(new4tcurf[,27]),mean=0,sd=elevhh)
 tmp4_rnorm(max(new4tcurf[,31]),mean=0,sd=elevcd)
 tmp5_rnorm(max(new4tcurf[,30]),mean=0,sd=elevwkld)
 tmpint_rnorm(max(new4tcurf[,29]),mean=0,sd=elevint)
 #add random effects to data
 for (i in 1:max(new4tcurf[,28])){
    new4tcurf[new4tcurf[,28]==i,35]_new4tcurf[new4tcurf[,28]==i,35]+tmp2[i]}
 for (i in 1:max(new4tcurf[,27])){
    new4tcurf[new4tcurf[,27]==i,35]_new4tcurf[new4tcurf[,27]==i,35]+tmp3[i]}
 for (i in 1:max(new4tcurf[,31])){
    new4tcurf[new4tcurf[,31]==i,35]_new4tcurf[new4tcurf[,31]==i,35]+tmp4[i]}
 for (i in 1:max(new4tcurf[,30])){
    new4tcurf[new4tcurf[,30]==i,35]_new4tcurf[new4tcurf[,30]==i,35]+tmp5[i]}
 for (i in 1:max(new4tcurf[,29])){
    new4tcurf[new4tcurf[,29]==i,35]_new4tcurf[new4tcurf[,29]==i,35]+tmpint[i]}
 new4tcurf[,35]_new4tcurf[,35]+rnorm(length(new4tcurf[,35]),mean=0,sd=elev1)
```

```
#add column names as first row of data
 new5tcurf_rbind(c(dimnames(newtcurf)[[2]],"normsimresp"),new4tcurf)
```

A.2 Normal Response Simulation Code: Single Month

The following S-Plus code was used to simulate a normal response variable onto the CURF structure for any single month with known random effect sizes at each of the levels. This is a small variation to the previous algorithm.

```
#choose month
 new4tcurf_new3tcurf[new3tcurf[,12]==1,]
#Sort, recode and resort levels for input to MLwiN
 new4tcurf_sort.col(new4tcurf,"@ALL",c("pindex"))
 plv1_cbind(unique(new4tcurf$pindex),1:length(unique(new4tcurf$pindex)))
 for (i in 1:length(plv1[,1])){
 new4tcurf[new4tcurf[,28]==plv1[i,1],28]_plv1[i,2]}
#Next households are in column 27
 new4tcurf_sort.col(new4tcurf,"@ALL",c("hhindex"))
 hhlv2_cbind(unique(new4tcurf$hhindex),1:length(unique(new4tcurf$hhindex)))
 for (i in 1:length(hhlv2[,1])){
 new4tcurf[new4tcurf[,27]==hhlv2[i,1],27]_hhlv2[i,2]}
#CDs are in column 31
 new4tcurf_sort.col(new4tcurf,"@ALL",c("cdindex"))
 cdlv3_cbind(unique(new4tcurf$cdindex),1:length(unique(new4tcurf$cdindex)))
 for (i in 1:length(cdlv3[,1])){
 new4tcurf[new4tcurf[,31]==cdlv3[i,1],31]_cdlv3[i,2]}
#Workloads are in column 30
 new4tcurf_sort.col(new4tcurf,"@ALL",c("wkldindex"))
 wklv5_cbind(unique(new4tcurf$wkldindex),1:length(unique(new4tcurf$wkldindex)))
 for (i in 1:length(wklv5[,1])){
 new4tcurf[new4tcurf[,30]==wklv5[i,1],30]_wklv5[i,2]}
#Workloads are in column 29
 new4tcurf_sort.col(new4tcurf,"@ALL",c("intindex"))
```

```
 intlv4_cbind(unique(new4tcurf$intindex),1:length(unique(new4tcurf$intindex)))
 for (i in 1:length(intlv4[,1])){
 new4tcurf[new4tcurf[,29]==intlv4[i,1],29]_intlv4[i,2]}
#Resort and simulate normal response with known random effects.
 new4tcurf_sort.col(new4tcurf,"@ALL",c("pindex","hhindex","cdindex","wkldindex"))
#set up effect sizes
 gmean_10
 elev2_sqrt(2)
 elevhh_sqrt(1)
 elevcd_sqrt(0.75)
 elevint_sqrt(0.2)
 elevwkld_0.5
 #data is new4tcurf
 #create column for output - 35
new4tcurf_cbind(new4tcurf[,1:34],rep(gmean,length(new4tcurf[,1])))
 #simulate random effect vectors
 tmp2_rnorm(max(new4tcurf[,28]),mean=0,sd=elev2)
 tmp3_rnorm(max(new4tcurf[,27]),mean=0,sd=elevhh)
 tmp4_rnorm(max(new4tcurf[,31]),mean=0,sd=elevcd)
 tmp5_rnorm(max(new4tcurf[,30]),mean=0,sd=elevwkld)
 tmpint_rnorm(max(new4tcurf[,29]),mean=0,sd=elevint)
 #add random effects to data
 for (i in 1:max(new4tcurf[,28])){
    new4tcurf[new4tcurf[,28]==i,35]_new4tcurf[new4tcurf[,28]==i,35]+tmp2[i]}
 for (i in 1:max(new4tcurf[,27])){
    new4tcurf[new4tcurf[,27]==i,35]_new4tcurf[new4tcurf[,27]==i,35]+tmp3[i]}
 for (i in 1:max(new4tcurf[,31])){
    new4tcurf[new4tcurf[,31]==i,35]_new4tcurf[new4tcurf[,31]==i,35]+tmp4[i]}
 for (i in 1:max(new4tcurf[,30])){
    new4tcurf[new4tcurf[,30]==i,35]_new4tcurf[new4tcurf[,30]==i,35]+tmp5[i]}
 for (i in 1:max(new4tcurf[,29])){
    new4tcurf[new4tcurf[,29]==i,35]_new4tcurf[new4tcurf[,29]==i,35]+tmpint[i]}
#add column names as first row of data
 new5tcurf_rbind(c(dimnames(newtcurf)[[2]],"normsimresp"),new4tcurf)
```

A.3 Binomial Response Simulation Code

To simulate a multilevel binary response vector, y, we can model the probability that a person's responses will be either in either state 0 or state 1 with the following multilevel logistic regression model.

$$\Pr(y_{ijlkmn} = 1 | \pi_{jklmn}) = \frac{\exp(\pi_{jklmn})}{1 + \exp(\pi_{jklmn})} \tag{A.1}$$

where π can be decomposed into a number of variance components corresponding to the classification levels in the CURF.

$$\pi_{jklmn} = \mu + \omega_n + \nu_{mn} + \varphi_{lmn} + \phi_{klmn} + \theta_{jlkmn}$$

and

- μ is a fixed effect
- The random effects are all normally distributed, i.e. $\omega_n \sim N(0, \sigma^2_\omega)$, $\nu_{mn} \sim N(0, \sigma^2_\nu)$, $\varphi_{lmn} \sim N(0, \sigma^2_\varphi)$, $\phi_{klmn} \sim N(0, \sigma^2_\phi)$ and $\theta_{jklmn} \sim N(0, \sigma^2_\theta)$
- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively

Hence, once these probabilities have been determined we can individually sample observations directly from a binomial distribution with the appropriate probabilities. The next S-Plus algorithm was used to simulate a binary response variable with known random effect sizes on the CURF structure.

```
#set up effect sizes
 gmean_2
 elev2_sqrt(22)
 elevhh_sqrt(35)
 elevcd_sqrt(5)
 elevint_sqrt(0.02)
 elevwkld_1
 #data is new3tcurf
```

```
 #create columns for output - 36,37
new4tcurf_cbind(new3tcurf,rep(gmean,length(new3tcurf[,1])),rep(0,length(new3tcurf[,1])))
 #Person index 28
 #hhindex 27
 #cdindex 31
 #Int index 29
 #Wkld index 30
 tmp2_rnorm(max(new4tcurf[,28]),mean=0,sd=elev2)
 tmp3_rnorm(max(new4tcurf[,27]),mean=0,sd=elevhh)
 tmp4_rnorm(max(new4tcurf[,31]),mean=0,sd=elevcd)
 tmp5_rnorm(max(new4tcurf[,30]),mean=0,sd=elevwkld)
 tmpint_rnorm(max(new4tcurf[,29]),mean=0,sd=elevint)
 for (i in 1:max(new4tcurf[,28])){
    new4tcurf[new4tcurf[,28]==i,36]_new4tcurf[new4tcurf[,28]==i,36]+tmp2[i]}
 for (i in 1:max(new4tcurf[,27])){
    new4tcurf[new4tcurf[,27]==i,36]_new4tcurf[new4tcurf[,27]==i,36]+tmp3[i]}
 for (i in 1:max(new4tcurf[,31])){
    new4tcurf[new4tcurf[,31]==i,36]_new4tcurf[new4tcurf[,31]==i,36]+tmp4[i]}
 for (i in 1:max(new4tcurf[,30])){
    new4tcurf[new4tcurf[,30]==i,36]_new4tcurf[new4tcurf[,30]==i,36]+tmp5[i]}
 for (i in 1:max(new4tcurf[,29])){
    new4tcurf[new4tcurf[,29]==i,36]_new4tcurf[new4tcurf[,29]==i,36]+tmpint[i]}
 new4tcurf_cbind(new4tcurf,exp(new4tcurf[,36])/(1+exp(new4tcurf[,36])),
    rep(1,length(new4tcurf[,1])),rep(1,length(new4tcurf[,1])))
 for (i in 1:length(new4tcurf[,1])){
    new4tcurf[i,37]_rbinom(1,1,new4tcurf[i,38])
    print(i)}
new5tcurf_rbind(c(dimnames(newtcurf)[[2]],"normsimresp","tempbin","binsimresp",
    "probbin"),new4tcurf)
```

A.4 Multinomial Response Simulation Code

In the multinomial case we choose a baseline category and model the probability that the response is any other state, i, in relation to this baseline. The following multilevel logistic regression model can be used to model the probability, $\mathrm{P}^{(\mathrm{i})}$, that

a person's responses, y, will be in one of t non-baseline states i compared with the baseline (state 0).

$$\mathrm{P^{(i)} = Pr(y_{ijlkmn} = i|\pi^{(i)}_{jklmn}) = \frac{exp(\pi^{(i)}_{jklmn})}{1 + \sum_{i=1}^{t} exp(\pi^{(i)}_{jklmn})}}$$

where $\pi^{(i)}$, can be decomposed into a number of variance components corresponding to the classification levels in the CURF.

$$\pi^{(i)}_{jklmn} = \mu^{(i)} + \omega^{(i)}_{n} + \nu^{(i)}_{mn} + \varphi^{(i)}_{lmn} + \phi^{(i)}_{klmn} + \theta^{(i)}_{jlkmn}$$

and

- The random effects are all normally distributed, i.e. $\omega^{(i)}_{n} \sim N(0, (\sigma^{(i)}_{\omega})^2)$, $\nu^{(i)}_{mn} \sim N(0, (\sigma^{(i)}_{\nu})^2)$, $\varphi^{(i)}_{lmn} \sim N(0, (\sigma^{(i)}_{\varphi})^2)$, $\phi^{(i)}_{klmn} \sim N(0, (\sigma^{(i)}_{\phi})^2)$ and $\theta^{(i)}_{jklmn} \sim N(0, (\sigma^{(i)}_{\theta})^2)$
- i, j, k, l, m and n are indices referring to the response/measurement level, the person/individual level, the dwelling/household level, the CD, interviewer and workload levels respectively

Then as we have specified all possible states, the probability of being in the baseline state can be simply obtained by subtraction.

$$\mathrm{P^{(0)} = 1 - \sum_{i=1}^{t} exp(\pi^{(i)}_{jklmn})}$$

Finally as we can now simulate the probabilities we can simulate an appropriate multinomial response using the following algorithm. First sample a random number, r_d drawn from the uniform (0,1) distribution. Then for all states t and the baseline state 0 generate the following response vector;

1. if $0 < r_d \leq \mathrm{P^{(0)}}$ then let $y_{sim} = 0$
2. if $\sum_{j=0}^{t-1} \mathrm{P^{(j)}} < r_d \leq \sum_{j=0}^{t} \mathrm{P^{(j)}}$ then let $y_{sim} = t$

The following S-Plus code was then used to simulate a multinomial response with known random effect sizes on the CURF structure.

```
 gmean_-5.746
 elev2_sqrt(43.34)
 elevhh_sqrt(7.136)
 elevcd_sqrt(7.433)
 elevint_sqrt(0.1483)
 elevwkld_sqrt(0.08344)
 gmean2_-4.763
 elev22_sqrt(120.8)
 elevhh2_sqrt(60.87)
 elevcd2_sqrt(19.33)
 elevint2_sqrt(0.1187)
 elevwkld2_sqrt(0.8267)
 tmp2_rnorm(max(new4tcurf[,28]),mean=0,sd=elev2)
 tmp3_rnorm(max(new4tcurf[,27]),mean=0,sd=elevhh)
 tmp4_rnorm(max(new4tcurf[,31]),mean=0,sd=elevcd)
 tmp5_rnorm(max(new4tcurf[,30]),mean=0,sd=elevwkld)
 tmp22_rnorm(max(new4tcurf[,28]),mean=0,sd=elev22)
 tmp32_rnorm(max(new4tcurf[,27]),mean=0,sd=elevhh2)
 tmp42_rnorm(max(new4tcurf[,31]),mean=0,sd=elevcd2)
 tmp52_rnorm(max(new4tcurf[,30]),mean=0,sd=elevwkld2)
 tmpint_rnorm(max(new4tcurf[,29]),mean=0,sd=elevint)
 tmpint2_rnorm(max(new4tcurf[,29]),mean=0,sd=elevint2)
#Cols 36,37 tmp cols for working out p0,p1,p2
 new4tcurf_cbind(new3tcurf,rep(gmean,length(new3tcurf[,1])),
    rep(gmean2,length(new3tcurf[,1])))
 for (i in 1:max(new4tcurf[,28])){
    new4tcurf[new4tcurf[,28]==i,36]_new4tcurf[new4tcurf[,28]==i,36]+tmp2[i]
    new4tcurf[new4tcurf[,28]==i,37]_new4tcurf[new4tcurf[,28]==i,37]+tmp22[i]}
 for (i in 1:max(new4tcurf[,27])){
    new4tcurf[new4tcurf[,27]==i,36]_new4tcurf[new4tcurf[,27]==i,36]+tmp3[i]
    new4tcurf[new4tcurf[,27]==i,37]_new4tcurf[new4tcurf[,27]==i,37]+tmp32[i]}
 for (i in 1:max(new4tcurf[,31])){
    new4tcurf[new4tcurf[,31]==i,36]_new4tcurf[new4tcurf[,31]==i,36]+tmp4[i]
    new4tcurf[new4tcurf[,31]==i,37]_new4tcurf[new4tcurf[,31]==i,37]+tmp42[i]}
 for (i in 1:max(new4tcurf[,30])){
    new4tcurf[new4tcurf[,30]==i,36]_new4tcurf[new4tcurf[,30]==i,36]+tmp5[i]
```

```
   new4tcurf[new4tcurf[,30]==i,37]_new4tcurf[new4tcurf[,30]==i,37]+tmp52[i]}
 for (i in 1:max(new4tcurf[,29])){
   new4tcurf[new4tcurf[,29]==i,36]_new4tcurf[new4tcurf[,29]==i,36]+tmpint[i]
   new4tcurf[new4tcurf[,29]==i,37]_new4tcurf[new4tcurf[,29]==i,37]+tmpint2[i]}
#Now add probabilities po,p1,p2 in final 3 columns. (+1 determ)
 new4tcurf_cbind(new4tcurf,rep(0,length(new4tcurf[,1])),rep(0,length(new4tcurf[,1])),
   rep(0,length(new4tcurf[,1])),runif(length(new4tcurf[,1]),0,1),
   rep(0,length(new4tcurf[,1])))
 new4tcurf[,39]_exp(new4tcurf[,36])/(1+exp(new4tcurf[,36])+exp(new4tcurf[,37]))
 new4tcurf[,40]_exp(new4tcurf[,37])/(1+exp(new4tcurf[,36])+exp(new4tcurf[,37]))
 new4tcurf[,38]_1-(new4tcurf[,39]+new4tcurf[,40])
#Set up rule for using determining prob
 for (i in 1:length(new4tcurf[,1])) {
   if (new4tcurf[i,38]<=new4tcurf[i,41]&new4tcurf[i,41]<=
       (new4tcurf[i,38]+new4tcurf[i,39])) {new4tcurf[i,42]_1}
   if ((new4tcurf[i,38]+new4tcurf[i,38])<=new4tcurf[i,41]&new4tcurf[i,41]<=1)
       {new4tcurf[i,42]_2}
   print(i)}
 new4tcurf[,42]_new4tcurf[,42]+1
 new5tcurf_rbind(c(dimnames(newtcurf)[[2]],"normsimresp","tmp1","tmp2","p0","p1","p2",
   "detrm","multsimresp"),new4tcurf)
 #new5tcurf[1,24]_"consl"
 #new5tcurf[1,33]_"bconsl"
 #new5tcurf[1,34]_"denoml"
 new6tcurf_new5tcurf[,c(2,3,4,5,6,7,8,24,26,27,28,29,30,31,33,34,42)]
 new6tcurf[1,]_paste(substring(substring(new6tcurf[1,],2,nchar(new6tcurf[1,])-1),1,4)
   ,"l",sep="")
 new6tcurf_new6tcurf[,c(17,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16)]
```

Appendix B

Matrix Partitioning Algorithm for $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$

The following S-plus code was used to calculate $\mathbf{F_{int}}$ and $\mathbf{F_{wk}}$ for the CURF. This algorithm partitions the **S** matrix into column vectors requiring less available RAM.

```
ptmp_t(za)%*%(abs((za%*%t(za)[,1])-(zb%*%t(zb)[,1])))
for (i in 2:length(za[,1])){
    ptmp_cbind(ptmp,t(za)%*%(abs((za%*%t(za)[,i])-(zb%*%t(zb)[,i]))))
    print(i)}
ptmp%*%za
```

Appendix C

MLwiN Macros

This section contains MLwiN macro code that was used to empirically estimate variances for interviewer effect estimates with logistic response variables. Loops cannot be nested within macros in MLwiN so a number of linked macro files were created for each task.

C.1 Empirical Estimate of the Variance of the Interviewer Effect Estimate with Logistic Response

There were three MLwiN macros created to perform this task. Setuplog.txt is a macro used to setup the model parameters and prepare the model for estimation. After changing the response to logistic using the MLwiN GUI and running IGLS to setup the starting values looplog.txt was run to repeatedly simulate and estimate the model parameters.

C.1.1 setuplog.txt Macro

```
note fpath c:\program files\mlwin2.0\mac
note generate vectors to start i, j, k initially full intergroup interpenetration
note i is individual = 1:2450
```

```
note j is interviewer = 1:70
note k is workload = 1:50
note here do 1 to 5 with step value 0.5 for the fixed effect parameter i.e. b1

seed 4 batch 1

note set up popn parameters
note b21 will be the fixed effect
note b22 will be the interviewer effect
note b23 will be the spatial effect

set b21 1
set b22 0.5
set b23 1
set b25 2450
set b26 70
set b27 50
calc b28=b25/b26
calc b29=b25/b27
generate 1 b25 1 C1
code b26 1 b28 C2
code b27 b29 1 c3
name c1 'Persn'
name c2 'Intvr'
name c3 'Wkld'

note now set values for fixed, random effects in c4,5,6 etc
put b25 b21 c4
name c4 'feff'
nran b26 c5
calc c6=c5*b22
merge c1 c6 c2 c7
name c7 'intresd'
nran b27 c8
calc c9=c8*b23
merge c1 c9 c3 c10
```

```
name c10 'wkldresd'
name c11 'piijk'
calc c11=c4+c10+c7
expo c11 c12
name c14 'probijk'
calc c14=c12/(1+c12)
bran b25 c15 c14 1
name c15 'binoresp'
name c16 'DENOM'
put b25 1 c16
put b25 1 c13
name c13 'cons'
resp c15
iden 3 c3
iden 2 c2
iden 1 c1
expl 1 c13
setv 3 c13
setv 2 c13
set b14 0
set b12 0
set b11 1

note easiest to initialize binomial response by hand. Then run once by
note MCMC before running separate macros.
```

C.1.2 looplog.txt Macro

```
note separate macro to loop and store data note will call rerunlog.txt
note to replace response col and run MCMC
note must run setuplog.txt first and start first run manually
note including setting up MCMC classification structure
note to test loop over fixed effect in b21

loop b21 2 5 1
```

```
note loop b36 1 2 1
fpath c:\program files\mlwin2.0\mac
obey rerunlog.txt

note store estimates for vc and fixed effect in cols

pick 1 c1096 b44
join c44 b44 c44
name c44 'vc3est'
pick 2 c1096 b44
join c45 b44 c45
name c45 'vc2est'
pick 3 c1096 b44
join c46 b44 c46
name c46 'resrest'
pick 1 c1098 b44
join c47 b44 c47
name c47 'fefest'
endloop

average c44 b52 b53 b54 b55
join c48 b54 c48
raise c48 2 c48
join c48 b53 c48
name c48 'sampvarv3'
average c45 b52 b53 b54 b55
join c49 b54 c49
raise c49 2 c49
join c49 b53 c49
name c49 'sampvarv2'
average c46 b52 b53 b54 b55
join c50 b54 c50
raise c50 2 c50
join c50 b53 c50
name c50 'sampvarr'
```

```
average c47 b52 b53 b54 b55
join c51 b54 c51
raise c51 2 c51
join c51 b53 c51
name c51 'sampvarf'

note output is in c18-c21 with sampling var above mean estimate
note to change length for samp var calculation
note erase c43
note pick 1 300 c45 c43 note aver c43
```

C.1.3 rerunlog.txt Macro

```
note as before loop and store data in separate macro that calls this one
set b121 b21
set b122 b22
set b123 b23
set b125 b25
set b126 b26
set b127 b27
calc b128=b125/b126
calc b129=b125/b127
generate 1 b125 1 C19
code b126 1 b128 C20
code b127 b129 1 c21
put b125 b121 c22
name c22 'feffb'
nran b126 c23
calc c24=c23*b122
merge c19 c24 c20 c25
name c25 'intresdb'
nran b127 c26
calc c27=c26*b123
merge c19 c27 c21 c28
name c28 'wkldresdb'
name c29 'piijkb'
```

```
calc c29=c22+c28+c25
expo c29 c30
name c32 'probijkb'
calc c32=c30/(1+c30)
bran b125 c33 c32 1
name c33 'binorespb'
name c34 'DENOMb'
put b125 1 c34
put b125 1 c31
name c31 'consb'
calc c15=c33

fpath c:\program files\mlwin2.0\discrete
mcrs 1
batch 0
link -5 g19 1
calc g18=g19
rlev 2
rfun
rcov 2
rout c1493 c1492
resi
join c1491 c1493 c1491
join c1490 c1492 c1490
erase c1493 c1492
rlev 3
rfun
rcov 2
rout c1493 c1492
resi
join c1491 c1493 c1491
join c1490 c1492 c1490
erase c1493 c1492
calc g19=g18
erase g18
link 0 g18
```

```
mcmc 0 500 1 5.8 50 10 c1491 c1490 2 2 1 1 1 2
erase c1090 c1091 c1491 c1490
mcmc 1 1000 1 c1090 c1091 c1003 c1004 1 2
pupn c1003 c1004
aver c1091 b99 b100
endobey
start
```

C.2 Macros to Estimate Variance of Interviewer Effect over Varying Degrees of Intergroup Interpenetration

These macros consider how the variance of the interviewer effect estimate changes as the degree of intergroup interpenetration changes. To run this set of macros apply setuplog.txt and initialize using the MLwiN GUI as before. Switch to MCMC estimation and run as a cross-classified model to prepare the starting values. Then run the firstloop.txt macro which will call looplog.txt and rerunlog.txt. As the output is stacked in a single column, resu.txt can be applied to sort the results appropriately.

C.2.1 setuplog2.txt Macro

```
note fpath c:\program files\mlwin2.0\mac
note generate vectors to start i, j, k initially full intergroup interpenetration
note i is individual = 1:5000
note j is interviewer = 1:100
note k is workload = 1:50

seed 4 batch 1

note set up popn parameters
note b21 will be the fixed effect
note b22 will be the interviewer effect
```

```
note b23 will be the spatial effect

set b21 2.5
set b22 1
set b23 2
set b25 5000
set b26 100
set b27 50
calc b28=b25/b26
calc b29=b25/b27
generate 1 b25 1 C1
code b26 1 b28 C2
code b27 b29 1 c3
name c1 'Persn'
name c2 'Intvr'
name c3 'Wkld'

note now set values for fixed, random effects in c4,5,6
put b25 b21 c4
name c4 'feff'
nran b26 c5
calc c6=c5*b22
merge c1 c6 c2 c7
name c7 'intresd'
nran b27 c8
calc c9=c8*b23
merge c1 c9 c3 c10
name c10 'wkldresd'
name c11 'piijk'
calc c11=c4+c10+c7
expo c11 c12
name c14 'probijk'
calc c14=c12/(1+c12)
bran b25 c15 c14 1
name c15 'binoresp'
name c16 'DENOM'
```

```
put b25 1 c16
put b25 1 c13
name c13 'cons'
resp c15
iden 3 c3
iden 2 c2
iden 1 c1
expl 1 c13
setv 3 c13
setv 2 c13
set b14 0
set b12 0
set b11 1

note easiest to initialize binomial response by hand. Then run once by
note MCMC before running separate macros.
```

C.2.2 firstloop.txt Macro

```
note this is just a macro file to loop over all required values
note of the degree of interpenetration
note for degrees of interpenetration which divide
note into the number of groups without remainder

loop b150 0.04 0.96 0.04

obey looplog2.txt
fpath c:\program files\mlwin2.0\mac\biglogloop
endloop
```

C.2.3 looplog2.txt Macro

```
note separate macro to loop and store data
note will call rerunlog2.txt to replace response col and run mcmc estimation
note must run setuplog2.txt first and start first run manually
note including setting up the mcmc classification structure
note to test loop over fixed effect in b21 to repeat loop over b36
```

```
note loop b21 2 5 1

loop b36 1 250 1
fpath c:\program files\mlwin2.0\mac\biglogloop
obey rerunlog2.txt

note store estimates for vc and fixed effect in cols
pick 1 c1096 b44
join c44 b44 c44
name c44 'vc3est'
pick 2 c1096 b44
join c45 b44 c45
name c45 'vc2est'
pick 3 c1096 b44
join c46 b44 c46
name c46 'resrest'
pick 1 c1098 b44
join c47 b44 c47
name c47 'fefest'
endloop

average c44 b52 b53 b54 b55
join c48 b54 c48
raise c48 2 c48
join c48 b53 c48
name c48 'sampvarv3'
average c45 b52 b53 b54 b55
join c49 b54 c49
raise c49 2 c49
join c49 b53 c49
name c49 'sampvarv2'
average c46 b52 b53 b54 b55
join c50 b54 c50
raise c50 2 c50
join c50 b53 c50
name c50 'sampvarr'
```

```
average c47 b52 b53 b54 b55
join c51 b54 c51
raise c51 2 c51
join c51 b53 c51
name c51 'sampvarf'
name c52 'dpen'
join c52 b150 c52
join c52 b150 c52

note output is in c18-c21 with sampling var above mean estimate
note to change length for samp var calculation
note erase c43
note pick 1 300 c45 c43 note aver c43
```

C.2.4 rerunlog2.txt Macro

```
note as before loop and store data in separate macro that calls this one
note set degree of interpenetration in overall loop firstloop.txt
note set b150 0.5

set b121 b21
set b122 b22
set b123 b23
set b125 b25
set b126 b26
set b127 b27
calc b128=b125/b126
calc b129=b125/b127
generate 1 b125 1 C19
code b126 1 b128 C20
code b127 b129 1 c21
put b125 b121 c22
name c22 'feffb'
nran b126 c23
calc c24=c23*b122
merge c19 c24 c20 c25
```

```
name c25 'intresdb'
nran b127 c26
calc c27=c26*b123
merge c19 c27 c21 c28
name c28 'wkldresdb'
name c29 'piijkb'
calc c29=c22+c28+c25
expo c29 c30
name c32 'probijkb'
calc c32=c30/(1+c30)
bran b125 c33 c32 1
name c33 'binorespb'
name c34 'DENOMb'
put b125 1 c34
put b125 1 c31
name c31 'consb'
calc c15=c33

note calc number of confounded workloads after initialization
calc b151=b126-b150*b127
round b151 b151
calc b152=b125/b151
round b152 b152

note set up new id vector to be joined
code b151 1 b152 C36 name c36 'newintid'

note now pick confounded bit first
calc b153=b150*b127
round b153 b153
calc c36=c36+b153
calc b154=b153*b129
pick 1 b154 c21 c35
name c35 'newwkid'
join c35 c36 c37
name c38 'newcompid'
```

```
pick 1 b125 c37 c38

note short code to work out new response vector
merge c19 c24 c38 c39
name c25 'intresdc'
name c40 'piijkc'
calc c40=c22+c28+c39
expo c40 c41
name c41 'probijkc'
calc c42=c41/(1+c41)
bran b125 c43 c42 1
name c43 'binorespc'

note generate data with new structure in following loops....
note swap back in new response and id vectors
calc c15=c43
calc c2=c38

fpath c:\program files\mlwin2.0\discrete
mcrs 1
batch 0
link -5 g19 1
calc g18=g19
rlev 2
rfun
rcov 2
rout c1493 c1492
resi
join c1491 c1493 c1491
join c1490 c1492 c1490
erase c1493 c1492
rlev 3
rfun
rcov 2
rout c1493 c1492
resi
```

```
join c1491 c1493 c1491
join c1490 c1492 c1490
erase c1493 c1492
calc g19=g18
erase g18
link 0 g18
mcmc 0 500 1 5.8 50 10 c1491 c1490 2 2 1 1 1 2
erase c1090 c1091 c1491 c1490
mcmc 1 2500 1 c1090 c1091 c1003 c1004 1 2
pupn c1003 c1004
aver c1091 b99 b100

endobey
start
```

C.2.5 resu.txt Macro

```
note stores count in cols
aver c47 b90
calc b91=b36-1
calc b92=b90/b91

note setup col for split
calc b100=100+b92
code b92 b91 1 c100
split c47 c100 c101-cb100
calc b103=100+b92
loop b93 1 b92 1
  calc b94 = b103+b93
  calc b95 = 100+b93
  aver cb95 b96 b97 b98 b99
  join cb94 b98 cb94
  raise cb94 2 cb94
  join cb94 b97 cb94
endloop
```

```
note now do the same thing for variance components c45 int c44 wk
calc b94=b94+2
calc b101=b94+b92
split c45 c100 cb94-cb101
loop b93 1 b92 1
  calc b102 = b101+b93
  calc b103 = b94+b93-1
  aver cb103 b96 b97 b98 b99
  join cb102 b98 cb102
  raise cb102 2 cb102
  join cb102 b97 cb102
endloop

note this bit works repeat for final vc
calc b102=b102+2
calc b101=b102+b92
split c44 c100 cb102-cb101
loop b93 1 b92 1
  calc b104 = b101+b93
  calc b103 = b102+b93-1
  aver cb103 b96 b97 b98 b99
  join cb104 b98 cb104
  raise cb104 2 cb104
  join cb104 b97 cb104
endloop
```

C.3 Empirical Exploration of Effect of the Sample Size on the Variance of the Interviewer Effect Estimate

This set of MLwiN macros can be used to consider how the variance of the interviewer effect estimate changes as the sample size changes, given that all other design and population parameters are held constant.

As before we initialize the macros by running setuplog3.txt first and then calling lpchgn.txt.

So in other words to get this to work we run setuplog.txt (after defining the appropriate fpath) then we use the MLwiN GUI to set the response variable as binomial and run. Then we choose MCMC estimation, set to cross-classified and run. Then run lpchgn.txt (note that we may have to reset fpath at this point)

Results from this run will be stacked in columns 48-52. Separate macros can be prepared to format these results appropriately.

C.3.1 setuplog3.txt Macro

```
seed 4
batch 1
note set up popn parameters
note b21 will be the fixed effect
note b22 will be the interviewer effect
note b23 will be the spatial effect
set b21 2.5
set b22 0.5
set b23 1.5
set b25 1500
set b26 100
set b27 50
calc b28=b25/b26
calc b29=b25/b27
generate 1 b25 1 C1
code b26 1 b28 C2
code b27 b29 1 c3
name c1 'Persn'
name c2 'Intvr'
name c3 'Wkld'

note now set values for fixed, random effects in c4,5,6
put b25 b21 c4
name c4 'feff'
```

```
nran b26 c5
calc c6=c5*b22
merge c1 c6 c2 c7
name c7 'intresd'
nran b27 c8
calc c9=c8*b23
merge c1 c9 c3 c10
name c10 'wkldresd'
name c11 'piijk'
calc c11=c4+c10+c7
expo c11 c12
name c14 'probijk'
calc c14=c12/(1+c12)
bran b25 c15 c14 1
name c15 'binoresp'
name c16 'DENOM'
put b25 1 c16
put b25 1 c13
name c13 'cons'
resp c15
iden 3 c3
iden 2 c2
iden 1 c1
expl 1 c13
setv 3 c13
setv 2 c13
set b14 0
set b12 0
set b11 1

note easiest to initialize binomial response by hand. Then run once by
note MCMC before running separate macros.
```

C.3.2 lpchgn.txt Macro

```
note macro to loop over changing sample size
```

```
note results will be stacked (2 rows each = estimate and samp var)
note in cols c48-c51

loop b25 500 1500 500
fpath c:\program files\mlwin2.0\mac\changen

note setup new data with increased n first
obey chgn.txt

note calculate sampling variance
obey looplog3.txt

note store sample sizes in col 52
join c52 b25 c52 join c52 b25 c52

endloop
```

C.3.3 chgn.txt Macro

```
note this file will change the sample size and loop over
note looplog macro to store the variance of the interviewer effect estimate
note start by setting up data and loop over b25 = sample size

calc b28=b25/b26
calc b29=b25/b27
generate 1 b25 1 C1
code b26 1 b28 C2
code b27 b29 1 c3
put b25 b21 c4
nran b26 c5
calc c6=c5*b22
merge c1 c6 c2 c7
nran b27 c8
calc c9=c8*b23
merge c1 c9 c3 c10
calc c11=c4+c10+c7
```

```
expo c11 c12
calc c14=c12/(1+c12)
bran b25 c15 c14 1
put b25 1 c16
put b25 1 c13
put b25 1 c17
```

C.3.4 looplog3.txt Macro

```
loop b36 1 4 1
fpath c:\program files\mlwin2.0\mac\changen
obey rerunlog3.txt
note store estimates for vc and fixed effect in cols
pick 1 c1096 b44
join c44 b44 c44
name c44 'vc3est'
pick 2 c1096 b44
join c45 b44 c45
name c45 'vc2est'
pick 3 c1096 b44
join c46 b44 c46
name c46 'resrest'
pick 1 c1098 b44
join c47 b44 c47
name c47 'fefest'
endloop

average c44 b52 b53 b54 b55
join c48 b54 c48
raise c48 2 c48
join c48 b53 c48
name c48 'sampvarv3'
average c45 b52 b53 b54 b55
join c49 b54 c49
raise c49 2 c49
join c49 b53 c49
```

```
name c49 'sampvarv2'
average c46 b52 b53 b54 b55
join c50 b54 c50
raise c50 2 c50
join c50 b53 c50
name c50 'sampvarr'
average c47 b52 b53 b54 b55
join c51 b54 c51
raise c51 2 c51
join c51 b53 c51
name c51 'sampvarf'
```

C.3.5 rerunlog3.txt Macro

```
set b121 b21
set b122 b22
set b123 b23
set b125 b25
set b126 b26
set b127 b27
calc b128=b125/b126
calc b129=b125/b127
generate 1 b125 1 C19
code b126 1 b128 C20
code b127 b129 1 c21
put b125 b121 c22
name c22 'feffb'
nran b126 c23
calc c24=c23*b122
merge c19 c24 c20 c25
name c25 'intresdb'
nran b127 c26
calc c27=c26*b123
merge c19 c27 c21 c28
name c28 'wkldresdb'
name c29 'piijkb'
```

```
calc c29=c22+c28+c25
expo c29 c30
name c32 'probijkb'
calc c32=c30/(1+c30)
bran b125 c33 c32 1
name c33 'binorespb'
name c34 'DENOMb'
put b125 1 c34
put b125 1 c31
name c31 'consb'
calc c15=c33

fpath c:\program files\mlwin2.0\discrete
mcrs 1
batch 0
link -5 g19 1
calc g18=g19
rlev 2
rfun
rcov 2
rout c1493 c1492
resi
join c1491 c1493 c1491
join c1490 c1492 c1490
erase c1493 c1492
rlev 3
rfun
rcov 2
rout c1493 c1492
resi
join c1491 c1493 c1491
join c1490 c1492 c1490
erase c1493 c1492
calc g19=g18
erase g18
link 0 g18
```

```
mcmc 0 500 1 5.8 50 10 c1491 c1490 2 2 1 1 1 2
erase c1090 c1091 c1491 c1490
mcmc 1 2500 1 c1090 c1091 c1003 c1004 1 2
pupn c1003 c1004
aver c1091 b99 b100

endobey
start
```

Glossary of Terms

ABS	Australian Bureau of Statistics
ACF	Autocorrelation Function
ANOVA	Analysis of Variance
ARA	Any Responsible Adult
B	Bias
CAR	Conditional Auto-Regressive
CD	Collection District
CSM	Covariance between Sampling and Measurement variance
CURF	Confidentialised Unit Record File
DEFF	Design Effect
DIC	Deviance Information Criterion
EM	Expectation Maximization algorithm
GIS	Geographical Information Systems
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GMM	Generalized Mixed Model
GUI	Graphical User Interface
HGLM	Hierarchical Generalized Linear Model
HH	Household
HLM	Hierarchical Linear Model
ICC	Intra-class Correlation Coefficient, also indicated by ρ
IGLS	Iterative Generalized Least Squares
LFS	Labour Force Survey
LMM	Linear Mixed Model

MAUP	Modifiable Areal Unit Problem
MCMC	Markov Chain Monte Carlo
MGA94	Map Grid of Australia 1994
MINQUE	Minimum Norm estimation
MINVAR	Minimum Variance estimation
MLE	Maximum Likelihood Estimation
MM	Multiple Membership
MMMC	Multiple Membership Multiple Classification
MSE	Mean Square Error
MPS	Monthly Population Survey
MQL	Marginal Quasi-Likelihood
MV	Measurement Variance
NA	Not Available, used to indicate missing values
NILF	Not In the Labour Force
OLS	Ordinary Least Squares
PACF	Partial Autocorrelation Function
PC	Personal Computer
PQL	Penalized Quasi-Likelihood
PSU	Primary Sampling Unit
REML	Restricted Maximum Likelihood
RIGLS	Restricted Iterative Generalized Least Squares
SAR	Simultaneous Auto-Regressive
SRSWR	Simple Random Sampling With Replacement
SRSWOR	Simple Random Sampling Without Replacement
SV	Sampling Variance
TV	Total Variance
VIF	Variance Inflation Factor
WLS	Weighted Least Squares

References

Afshartous, D. (1995) Determination of sample size for multilevel model design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Aitken, M., Anderson, D. A. and Hinde, J. (1981) Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, A*, **144**, 419–461.

Anderson, D. A. and Aitken, M. (1985) Variance component models with binary response: interviewer's variability. *Journal of the Royal Statistical Society, B*, **47**, 203–210.

Bailar, B. (1968) Recent research in reinterview procedures. *Journal of the American Statistical Association*, **63**, 41–63.

— (1976) Some sources of error and their effect on census statistics. *Demography*, **13**, 273–286.

Bailar, B. and Dalenius, T. (1969) Estimating the response variance components of the us bureau of the census model. *The Indian Journal of Statistics*, **31**, 341–360.

Bao, S. (2001) Literature review of spatial statistics and models. Paper available online at http://141.211.136.209/cdc/docs/review.pdf.

Barr, A. (1957) Differences between experienced interviewers. *Applied Statistics*, **6**, 180–188.

Bassi, F. and Fabbris, L. (1997) Estimators of nonsampling errors in interview-reinterview supervised surveys with interpenetrated assignments. In *Survey Measurement and Process Quality* (eds. L. Lyberg, P. Beimer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), chap. 32, 733–751. New York: John Wiley and Sons.

Beerten, R. (1999) The effect of interviewer and area characteristics on survey response rates: an exploratory analysis. *Tech. Rep. 45*, Office for National Statistics, London. Survey Methodology Bulletin.

Beimer, P., Groves, R., Lyberg, L., Mathiowetz, N. and Sudman, S. (eds.) (1991) *Measurement Errors in Surveys*. New York: John Wiley and Sons.

Beimer, P. and Lyberg, L. (eds.) (2003) *Introduction to Survey Quality.* New York: John Wiley and Sons.

Biemer, P. (1978) *The Estimation of Nonsampling Variance Components in Sample Surveys*. Ph.D. thesis, Texas A and M University.

Biemer, P. and Stokes, S. (1985) Optimal design of interviewer variance experiment in complex surveys. *Journal of the American Statistical Association*, **80**, 158–166.

Bivand, R. (1998) A review of spatial statistical techniques for location studies. Paper available online at www.nhh.no/geo/gib/gib1998/gib98-3/lund.html.

Bowley, A. (1915) *The Nature and Purpose of the Measurement of Social Phenomena.* London: P.S. King and Son, Ltd.

— (1926) Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, **22**, 1–62. Special annex following page 451.

Brackstone, G. (1999) Managing data quality in a statistical agency. *Survey Methodology*, **25**, 139–149.

Breslow, N. (2003) Whither pql? UW Biostatistics Working Paper Series.

Breslow, N. and Clayton, D. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brewer, K. (1963) A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, **5**, 5–13.

— (1979) A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, **74**, 911–915.

Brooks, S. and Roberts, G. (1998) Assessing convergence of markov chain monte carlo algorithms. *Statistics and Computing*, **8**, 319–335.

Browne, W. (1998) *Applying MCMC Methods to Multi-level Models.* Ph.D. thesis, University of Bath.

— (2002) *MCMC Estimation in MLwiN*. Institute of Education, University of London, London.

— (2004) An illustration of the use of reparameterisation methods for improving mcmc efficiency in crossed random effect models. *Multilevel Modelling Newsletter*, **16**, 13–25.

Browne, W. and Draper, D. (2000) Implementation and performance issues in the bayesian and likelihood fitting of multilevel models. *Computational Statistics*, **15**, 391–420.

Browne, W., Goldstein, H. and Rasbash, R. (2001) Multiple membership multiple classification (mmmc) models. *Statistical Modelling*, **1**, 103–124.

Browne, W. J. and Draper, D. (2003) A comparison of bayesian and likelihood-based methods for fitting multilevel models. Paper available online at http://www.maths.nott.ac.uk/personal/pmzwjb/materials/wbrssa.pdf.

Byrk, A. and Raudenbush, S. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage.

Campanelli, P., Sturgis, P. and Purdon, S. (1997) Can you hear me knocking? an investigation into the impact of interviwers on survey response rates. *Tech. rep.*, Social and Community Planning Research, London.

Case, A. (1991) Spatial patterns in household demand. *Econometrica*, **59**, 953–965.

Cochran, W. (1977) *Sampling Techniques*. New York: Wiley, 3 edn.

Cohen, M. (1998) Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, **14**, 267–275.

Collins, M. (1980) Interviewer variability. *Journal of the Market Research Society*, **22**, 77–95.

Cowles, M. and Carlin, B. (1996) Markov chain monte carlo convergence diagnostics. a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Deming, W. (1944) On errors in surveys. *American Sociological Review*, **9**, 359–369.

— (1950) *Some Theory of Sampling*. New York: John Wiley and Sons.

— (1960) *Sample Design in Business Research*. New York: John Wiley and Sons.

Ecob, R. and Jamieson, B. (1992) A multilevel analysis of interviewer effects on a health survey. In *Survey and Statisical Computing* (eds. A. Westlake, R. Banks, C. Payne and T. Orchard), 255–268. Elsevier Science Publishers B.V.

Feldman, J., Hyman, H. and Hart, C. (1951) A field study of interviewer effects on the quality of survey data. *Public Opinion Quarterly*, **15**, 734–761.

Felligi, I. (1964) Response variance and its estimation. *Journal of the American Statistical Association*, **59**, 1016–1041.

— (1974) An improved method of estimating the correlated response variance. *Journal of the American Statistical Association*, **69**, 496–501.

Forsman, G. (1989) Early survey models and their use in survey quality work. *Journal of Official Statistics*, **5**, 41–55.

Freeman, J. and Butler, E. (1976) Some sources of interviewer variance in surveys. *Public Opinion Quarterly*, **40**, 79–91.

Gelfand, A., Sahu, S. and Carlin, B. (1995) Efficient parameterizations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Gelfand, A. and Smith, A. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 479–488.

Gelman, A. and Rubin, D. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Geman, S. and Geman, D. (1984) Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 721–741.

Geweke, J. (1992) Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (eds. J. Bernado, J. Berger, A. Dawid and A. Smith), 169–193. Oxford: Oxford University Press.

Gilks, W., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gilks, W. and Wild, P. (1992) Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society C*, **41**, 337–348.

Gilmore, A., Anderson, R. and Rae, A. (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593–599.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

— (1987) Multilevel covariance component models. *Biometrika*, **74**, 430–431.

— (1989) Restricted unbiased iterative generalized least squares estimation. *Biometrika*, **76**, 622–623.

— (1991) Multilevel modelling of survey data. *The Statistician*, **40**, 235–244.

— (1995) *Multilevel Statistical Models*. London: Edward Arnold.

Goldstein, H. and Rasbash, J. (1992) Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics and Data Analysis*, **13**, 63–71.

— (1996) Improved approximations for multilevel models with binary response. *Journal of the Royal Statistical Society, A*, **159**, 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M. (1998) *A users guide to MLwiN*. Institute of Education, London.

Groves, R. (1989) *Survey Errors and Survey Costs*. New York: John Wiley and Sons.

Groves, R., Cialdini, R. and Couper, M. (1992) Understanding the decision to participate in a survey. *Public Opinion Quarterly*, **56**, 475–495.

Groves, R. and Couper, M. (1998) *Non response in household interview surveys.* New York: John Wiley and Sons.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. and Tourangeau, R. (2004) *Survey Methodology.* New York: John Wiley and Sons.

Groves, R. and Magilavy, L. (1986) Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, **50**, 251–266.

Hansen, M. and Hurwitz, W. (1943) On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, **14**, 333–362.

Hansen, M., Hurwitz, W. and Bershad, M. (1961) Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, **38**, 359–374.

Hansen, M., Hurwitz, W. and Madow, W. (1953) *Sample Survey Methods and Theory.* New York: John Wiley and Sons. Volume 1 Methods and applications. Volume 2 Theory.

Hansen, M., Hurwitz, W., Marks, E. and Maudlin, W. (1951) Response errors in surveys. *Journal of the American Statistical Association*, **46**, 147–190.

Hansen, M., Hurwitz, W. and Pritzker, L. (1964) The estimation and interpenetration of gross differences and the simple response variance. In *Contributions to Statistics, Presented to Professor P.C. Mahalanobis on the Occasion of his 70th Birthday* (eds. C. Rao, D. Lahiri, K. Nair, P. Pant and S. Shrikhande), 111–136. Oxford England: Statistical Publishing Society.

— (1967) Standardization of procedures for the evaluation of data. measurement errors and statistical standards in the bureau of the census. *Bulletin of the International Statistical Institute*, **42**, 49–66.

Hansen, M., Madow, W. and Teping, B. (1983) An evaluation of model dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, **78**, 776–793.

Hansen, M. and Marks, E. (1958) Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, **53**, 635–655.

Hartley, H. and Rao, J. (1978) The estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement* (ed. N. Krishnan Namboodiri). New York: Academic Press.

Haslett, S. (2003) Estimability and connectedness in generalized linear models, including links to small area estimation. Unpublished presentation notes.

Hastings, W. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.

Hill, S. and Smith, A. (1992) Parametrization issues in bayesian inference. In *Bayesian Statistics 4* (eds. J. Bernado, J. Berger, A. Dawid and A. Smith), 227–246. Oxford: Oxford University Press.

Hobert, J. and Casella, G. (1996) The effect of improper priors on gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**, 1461–1473.

Hox, J., de Leeuw, E. and Kreft, I. (1991) The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys* (eds. P. Beimer, R. Groves, L. Lyberg, N. Mathiowetz and S. Sudman), chap. 22, 439–461. New York: John Wiley and Sons.

Hox, J. J. (1994) Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, **28**, 300–318.

Hutchison, D. and Healy, M. (2001) The effect on variance component estimates of ignoring a level in a multilevel model. *Multilevel Modelling Newsletter*, **13**, 4–5.

Jones, K. (1992) Using multilevel models for survey analysis. In *Survey and Statisical Computing* (eds. A. Westlake, R. Banks, C. Payne and T. Orchard), 231–242. Elsevier Science Publishers B.V.

Kass, R., Tierney, L. and Kadane, J. (1988) Asymptotics in bayesian computation. In *Bayesian Statistics 3* (eds. J. Bernado, M. DeGroot, D. Lindley and A. Smith), 261–278. Oxford: Oxford University Press.

Kiaer, A. (1897) *The Representative Method of Statistical Surveys*. Oslo: Kristiania. 1976 translation from original Norwegian.

Kish, L. (1962) Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, **57**, 92–115.

— (1965) *Survey Sampling*. New York: John Wiley and Sons.

Koch, G. (1973) An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means. *Journal of the American Statistical Association*, **68**, 906–913.

Koop, J. (1974) Notes for a unified theory of estimation for sample surveys taking into account response errors. *Metrika*, **28**, 19–39.

Kruskal, W. (1991) Introduction. In *Measurement Errors in Surveys* (eds. P. Beimer, R. Groves, L. Lyberg, N. Mathiowetz and S. Sudman), xxiii–xxxiii. New York: John Wiley and Sons.

Kuk, A. (1995) Asymptotically unbiased estimation in generalized linear mixed models with random effects. *Journal of the Royal Statistical Society, B*, **57**, 395–407.

Kuk, A. and Cheng, Y. (1999) Pointwise and functional approximations in monte carlo maximum likelihood estimation. *Statistics and Computing*, **9**, 91–99.

Lee, Y. and Nelder, J. (1996) Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, B*, **58**, 619–678.

de Leeuw, J. and Kreft, I. (2001) Software for multilevel analysis. In *Multilevel Modelling of Health Statistics* (eds. A. Leyland and H. Goldstein), chap. 13, 187–204. Chichester: John Wiley and Sons.

Lessler, J. (1976) Survey designs which employ double sampling schemes for eliminating measurement process bias. In *American Statistical Association Proceedings of the Social Statistics Section*, vol. 2, 520–525. American Statistical Association.

— (1977) Use of error rates and discrepancy rates to calculate the mean square error of survey data. Paper presented at the annual meeting of the American Public Health Association.

— (1983) An expanded survey error model. In *Incomplete Data in Sample Surveys* (eds. W. Madow and I. Olkin), 259–270. New York: Academic Press.

Lessler, J. and Kalsbeek, W. (1992) *Non-sampling error in surveys.* New York: John Wiley and Sons.

Littell, R., Milliken, G., Stroup, W. and Wolfinger, R. (1996) *SAS System for Mixed Models.* SAS Institute Inc, Cary, NC.

Lohr, S. L. (1999) *Sampling: Design and Analysis.* Boston: Duxbury Press.

Longford, N. (1993) *Random Coefficient Models.* Oxford Science Publications. Oxford.

Lyberg, L., Beimer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. (1997a) Preface. In *Survey Measurement and Process Quality* (eds. L. Lyberg, P. Beimer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), xiii–xvii. New York: John Wiley and Sons.

Lyberg, L., Beimer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. (eds.) (1997b) *Survey Measurement and Process Quality.* New York: John Wiley and Sons.

MacEachern, S. and Berliner, L. (1994) Subsampling the gibbs sampler. *The American Statistician*, **48**, 188–190.

Mahalanobis, P. (1946) Recent experiments in statistical sampling in the indian statistical institute. *Journal of the Royal Statistical Society*, **109**, 325–370.

Martin, J. and Beerten, R. (2002) The effect of interviewer characteristics on survey response rates. Available on the web. Office of National Statistics.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*. London: Chapman and Hall, 2 edn.

McCulloch, C. and Searle, S. (2001) *Generalized, Linear, and Mixed Models*. New York: John Wiley and Sons.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Moerbeek, M., Gerard, J. and Berger, M. (2000) Design issues for experiments in multilevel populations. *Journal of Educational and Behavioural Statistics*, **25**, 271–284.

— (2001a) Optimal experimental designs for multilevel logistic models. *The Statistician*, **50**, 17–30.

— (2001b) Optimal experimental designs for multilevel models with covariates. *Communications in Statistics - Theory and Methods*, **30**, 2683–2698.

Moerbeek, M. and Wong, W. (2002) Multiple-objective optimal designs for the hierarchical linear model. *Journal of Official Statistics*, **18**, 291–303.

Mok, M. (1995) Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter*, **7**, 11–15.

Moser, C. and Kalton, G. (1972) *Survey Methods in Social Investigation*. New York: Basic Books Inc.

Murthy, M. (1967) *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–606.

Noh, M. and Lee, Y. (2004) Review of estimating methods for binary data in generalised linear mixed models. Submitted.

Normand, S. and Zou, K. (2002) Sample size considerations in observational health care quality studies. *Statistics in Medicine*, **21**, 331–345.

O'Muirceartaigh, C. and Wiggins, R. (1981) The impact of interviewer variability in an epidemiological survey. *Psychological Medicine*, **11**, 817–824.

O'Muircheartaigh, C. (1977) Statistical analysis in the context of survey research. In *The Analysis of Survey Data* (eds. C. O'Muircheartaigh and C. Payne), 193–239. New York: John Wiley and Sons.

— (1997) Measurement error in surveys: A historical perspective. In *Survey Measurement and Process Quality* (eds. L. Lyberg, P. Beimer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), 1–25. New York: John Wiley and Sons.

O'Muircheartaigh, C. and Campanelli, P. (1998) The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society A*, **161**, 63–77.

— (1999) A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society A*, **162**, 437–446.

Pannekoek, J. (1991) A mixed model for analyzing measurement errors for dichotomous variables. In *Measurement Errors in Surveys* (eds. P. Beimer, R. Groves, L. Lyberg, N. Mathiowetz and S. Sudman), chap. 25, 517–530. New York: John Wiley and Sons.

Pfeffermann, D., Kreiger, A. M. and Rinott, Y. (1998) Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, 1087–1114.

Pickery, J. (2000) *Applications of Multilevel Analysis in Survey Data Quality Research.* Ph.D. thesis, Katholieke Univesiteit Leuven, Faculteit Sociale Wetenschappen. Departement Sociologie, K.U.Leuven, E. van Evenstraat 2B, B 3000 Leuven, Belgie.

Pickery, J. and Loosveldt, G. (2000) Modeling interviewer effects in panel surveys: an application. *Survey Methodology*, **26**, 189–198.

— (2001) An exploration of question characteristics that mediate interviewer effects on item nonresponse. *Journal of Official Statistics*, **17**, 337–350.

— (2004) A simultaneous analysis of interviewer effects on various data quality indicators with identification of exceptional interviewers. *Journal of Official Statistics*, **20**, 77–89.

Platek, R. and Gray, G. (1983) Imputation methodology, total survey error, part v. In *Incomplete Data in Sample Surveys* (eds. W. Madow and I. Olkin), 249–333. New York: Academic.

Platek, R. and Sarndal, C. (2001) Can a statistician deliver? *Journal of Official Statistics*, **17**, 1–20.

Raferty, A. and Lewis, S. (1992) How many iterations in the gibbs sampler. In *Bayesian Statistics 4* (eds. J. Bernado, J. Berger, A. Dawid and A. Smith), 763–773. Oxford: Oxford University Press.

Raj, D. (1968) *Sampling Theory.* New York: McGraw Hill.

Rasbash, J. and Browne, W. (2001) Modelling non-hierarchical structures. In *Mutlilevel Modelling of Health Statistics* (eds. A. Leyland and H. Goldstein), 93–105. New York: John Wiley and Sons.

Rasbash, J. and Goldstein, H. (1994) Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics*, **19**, 337–350.

Rasbash, J., Steel, F. and Browne, W. (2003) *A User's Guide to MLwiN Version 2.0*. Institute of Education, London, University of London, 2.1 edn.

Raudenbush, S. (1993) A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, **18**, 321–349.

Rice, S. (1929) Contagious bias in the interview. *American Journal of Sociology*, **35**, 420–423.

Richardson, S. and Monfort, C. (2000) Ecological correlation studies. In *Spatial Epidemiology: Methods and Applications* (eds. P. Elliott, J. Wakefield, N. Best and D. Briggs), chap. 11, 205–220. Oxford: Oxford University Press.

Roberts, G. (1996) Markov chain concepts relating to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (eds. W. Gilks, S. Richardson and D. Spiegelhalter), 45–57. London: Chapman and Hall.

Rodriguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society A*, **158**, 73–89.

— (2001) Improved estimation procedures for multilevel models with binary response: a case study. *Journal of the Royal Statistical Society*, **164**, 339–355.

Royall, R. and Herson, J. (1973) Robust estimation in finite populations. *Journal of the American Statistical Association*, **68**, 880–893.

Sarndal, C., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sastry, N., Ghosh-Dastidar, B., Adams, J. and Pebley, A. (2003) The design of a multilevel study of children, families, and communities: The los angeles family and neighborhood survey. Office of Population Research Princeton University Working Paper Series No. 2003-06.

Schwarz, N. and Sudman, S. (1996) *Answering Questions: Methodology for determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey-Bass.

Searle, S. (1971) *Linear Models*. New York: John Wiley and Sons.

Searle, S., Casella, G. and McCulloch, C. (1992) *Variance Components*. New York: John Wiley and Sons.

Skinner, C., Holt, D. and Smith, D. (1989) *Analysis of Complex Surveys*. Chichester: John Wiley.

Smith, T. (1994) Sample survey 1975 to 1990; an age of reconciliation? *International Statistical Review*, **62**, 5–34.

Snijders, T. (1996) Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, **30**, 405–426.

Snijders, T. and Bosker, R. (1993) Standard errors and sample sizes for two level research. *Journal of Educational Statistics*, **18**, 237–259.

— (1999) *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1994) *BUGS Bayesian Inference using Gibbs Sampling. Version 0.30a*. MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. Best, N., Carlin, B. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, **64**, 583–640.

Sudman, S. and Bradburn, N. (1974) *Response Effects in Surveys*. Chicago: Aldine.

Sudman, S., Bradburn, N. and Schwarz, N. (1996) *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Sukhatme, P. and Seth, G. (1952) Nonsampling errors in surveys. *Journal of Indian Society of Agricultural Statistics*, **5**, 5–41.

Sukhatme, P. and Sukhatme, B. (1970) *Sampling Theory of Surveys with Applications*. Ames, IA: Iowa State University Press, 2 edn.

Tagashira, N. and Okabe, A. (2002) The modifiable areal unit problem in a regression model whose independent variable is a distance from a predetermined point. *Geographical Analysis*, **34**, 1–20.

Thompson, M. (1997) *Theory of Sample Surveys*. London: Chapman and Hall.

Tierney, L. (1996) Introduction to general state space markov chain theory. In *Markov Chain Monte Carlo in Practice* (eds. W. Gilks, S. Richardson and D. Spiegelhalter), 59–74. London: Chapman and Hall.

Tranmer, M. and Steel, D. (2001) Ignoring a level in a multilevel model: evidence from uk census data. *Environment and Planning*, **33**, 941–948.

Valliant, R., Dorfman, A. and Royall, R. (2000) *Finite Population Sampling and Inference*. Wiley Series in Probability and Statistics. New York: Wiley.

Verbeke, G. and Molenburghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. New York: Springer.

Waddington, D. and Thompson, R. (2004) Using a correlated probit model approximation to estimate the variance for binary matched pairs. *Statistics and Computing*, **14**, 83–90.

Wakefield, J., Best, N. and Waller, L. (2000) Bayesian approaches to disease mapping. In *Spatial Epidemiology: Methods and Applications* (eds. P. Elliott, J. Wakefield, N. Best and D. Briggs), chap. 11, 205–220. Oxford: Oxford University Press.

Wiggins, R., Longford, N. and O'Muircheartaigh, C. (1992) A variance components approach to interviewer effects. In *Survey and Statisical Computing* (eds. A. Westlake, R. Banks, C. Payne and T. Orchard), 243–254. Elsevier Science Publishers B.V.

Yamane, T. (1967) *Elementary Sampling Theory*. Englewood Cliffs, N.J.: Prentice Hall.

Yang, M. and Goldstein, H. (1996) Multilevel models for longitudinal data. In *Analysis of Change. Advanced Techniques in Panel Data Analysis* (eds. U. Engel and J. Reinecke), 191–220. Berlin: Walter de Gruyter and Co.

Zarkovich, S. (1966) Quality of stastical data. Food and Agricultural Organization of the United Nations.